COMPLETED
120

Y3.N88:25/4496

NUREG/CR-4496

A System for Generating Long Streamflow Records for Study of Floods of Long Return Period

Prepared by R. K. Linsley, B. A. Kraeger, D. D. Franz

Linsley, Kraeger Associates LTD.

Prepared for
U.S. Nuclear Regulatory
Commission

BEST COPY AVAILABLE

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability of responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

NOTICE

Availability of Reference Materials Cited in NRC Publications

Most documents cited in NRC publications will be available from one of the following sources:

1. The NRC Public Document Room, 1717 H Street, N.W. Washington, DC 20555
2. The Superintendent of Documents, U.S. Government Printing Office, Post Office Box 37082, Washington, DC 20013-7082
3. The National Technical Information Service, Springfield, VA 22161

Although the listing that follows represents the majority of documents cited in NRC publications, it is not intended to be exhaustive.

Referenced documents available for inspection and copying for a fee from the NRC Public Document Room include NRC correspondence and internal NRC memoranda; NRC Office of Inspection and Enforcement bulletins, circulars, information notices, inspection and investigation notices; Licensee Event Reports; vendor reports and correspondence; Commission papers; and applicant and licensee documents and correspondence.

The following documents in the NUREG series are available for purchase from the GPO Sales Program: formal NRC staff and contractor reports, NRC-sponsored conference proceedings, and NRC booklets and brochures. Also available are Regulatory Guides, NRC regulations in the *Code of Federal Regulations*, and *Nuclear Regulatory Commission Issuances*.

Documents available from the National Technical Information Service include NUREG-series reports and technical reports prepared by other federal agencies and reports prepared by the Atomic Energy Commission, forerunner agency to the Nuclear Regulatory Commission.

Documents available from public and special technical libraries include all open literature items, such as books, journal and periodical articles, and transactions. *Federal Register* notices, federal and state legislation, and congressional reports can usually be obtained from these libraries.

Documents such as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings are available for purchase from the organization sponsoring the publication cited.

Single copies of NRC draft reports are available free, to the extent of supply, upon written request to the Division of Technical Information and Document Control, U.S. Nuclear Regulatory Commission, Washington, DC 20555.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at the NRC Library, 7920 Norfolk Avenue, Bethesda, Maryland, and are available there for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from the American National Standards Institute, 1430 Broadway, New York, NY 10018.

BEST COPY AVAILABLE

NUREG/CR-4496
RG

A System for Generating Long Streamflow Records for Study of Floods of Long Return Period

Manuscript Completed: April 1985
Date Published: February 1986

R. K. Linsley, B. A. Kraeger, D. D. Franz

Linsley, Kraeger Associates LTD.
527 Bayview Drive
Aptos, CA 95003

Prepared for
Division of Risk Analysis and Operations
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555
NRC FIN D1562
Under Contract No. NRC 04-84-114

I

BEST COPY AVAILABLE

TABLE OF CONTENTS

List of Tables iii

List of Figures iv

I. INTRODUCTION 1

II. THE CONCEPT 3

III. THE HYDROLOGIC SIMULATION MODEL. 6

IV. THE STUDY CATCHMENT 8

V. THE STOCHASTIC RAINFALL 23

Related work. 24

Description of the Model. 27

Single Station Model. 28

Estimation of Parameters 30

Multivariate Model 42

Testing the Model 48

Discussion of Results 58

VI. THE HYDROLOGIC SIMULATION 63

VII. CONCLUSIONS 82

VIII. FOLLOW-ON STUDIES RESEARCH 84

REFERENCES 86

APPENDIX A. Mathematical Topics 89

Multivariate Normal Distribution . . . 89

Sampling from the Multivariate distribution 90

Generation of Random Numbers 91

BEST COPY AVAILABLE

LIST OF TABLES

4-1. Characteristics of the Study Basin 8

4-2. Catchment parameters for the Russian R. above Cloverdale 11

4-3. Comparison of observed and simulated annual flow volume, Russian R. nr Cloverdale, CA 22

5-1. Truncation Level and Lag-1 Correlation at Redwood Valley 29

5-2. Storm Length Distribution Results at Yorkville . . . 37

5-3. Storm Length Distribution in Season 5 at Yorkville . 38

5-4. Transformation Parameters for Hourly Rainfall at Yorkville 42

5-5. Correlation Matrices for Season 6 44

5-6. Summary of Correlograms for Season 6 46

5-7. Tests of Explicitly Mimicked Characteristics 53

5-8 Storm Length Distribution Comparisons 54

5-9. Hourly Rainfall Amount Distribution Comparison . . . 55

5-10. Interstorm Length Comparison 56

5-11. Interstorm Length Distribution for Season 6 Redwood Valley 57

BEST COPY AVAILABLE

LIST OF FIGURES

2-1. Location of test basins for the concept 4

3-1. Flow diagram for SWM84 7

4-1. The Russian River above Cloverdale showing reaches and station locations 9

4-2. Map showing the location of the study basin in California 10

4-3. Comparison of observed and simulated flood probabilities, Russian River nr Cloverdale, CA 1952 - 72 12

4-4 Plot of simulated and observed mean daily flows at Cloverdale for the water year 1956 13

4-5. Plot of simulated and observed mean daily flows at Cloverdale for the water year 1958 14

4-6. Plot of simulated and observed mean daily flows at Cloverdale for the water year 1965 15

4-7. Observed and simulated hydrographs for the flood of December, 1955 at Cloverdale 17

4-8. Observed and simulated hydrographs for the flood of January, 1956 at Cloverdale 18

4-9. Observed and simulated hydrographs for the flood of February, 1960 at Cloverdale 19

4-10. Observed and simulated hydrographs for the flood of December, 1964 at Cloverdale 20

4-11. Observed and simulated flow duration curves for the Russian River near Cloverdale, 1952-72 . . . 21

5-1. The assumed background process for rainfall 29

5-2. Scatter plot of mean hourly rainfall during storms . 51

6-1. Frequency curves for the 1000-yr and "observed" simulated samples using the Extreme Value type I distribution. The log Pearson curve is plotted for comparison. 64

BEST COPY AVAILABLE

6-2. Frequency curves from samples of 20, 50, 100, 500 and 1000 years drawn from the 1000-yr synthetic record 66

6-3. Comparison of flow duration curves for the Russian River nr. Cloverdale based on the 1000-yr sample and the simulated historic record 67

6-4. Comparison of frequency of annual flows at Cloverdale for the 1000-yr sample and the 20-yr simulated historic record 68

6-5. Comparison of the frequency of annual rainfalls at Redwood Valley for the 1000-yr stochastic generation and the historic record 69

6-6. Comparison of frequency of annual rainfalls at Yorkville from the 1000-yr stochastic generation and the historic record 70

6-7. Range of mean flood estimates at Cloverdale as a function of sample size 71

6-8. Range of standard deviation of annual floods at Cloverdale as a function of sample size 72

6-9. Range of skew of annual floods at Cloverdale as a function of sample size 73

6-10. Range of mean annual runoff at Cloverdale as a function of sample size 74

6-11. Range of standard deviation of annual runoff at Cloverdale as a function of sample size 75

6-12. Range of mean annual rainfall at Yorkville as a function of sample size 77

6-13. Range of standard deviation of rainfall at Yorkville and Redwood Valley as a function of sample size 78

6-14. Comparison of rainfall intensity and probability for various durations at Yorkville from observed and and synthetic 1000-yr samples 79

6-15. Comparison of rainfall intensity and probability for various durations at Redwood Valley from observed and synthetic 1000-yr samples 80

6-16. Probable Maximum Flood, Russian R. nr Cloverdale. . . 81

BEST COPY AVAILABLE

A SYSTEM FOR GENERATING LONG STREAMFLOW RECORDS FOR STUDY OF FLOODS OF LONG RETURN PERIOD

I. INTRODUCTION

The Nuclear Regulatory Commission (NRC) is responsible for assuring the safety of nuclear power plants. Since cooling water is essential for the efficient operation of any thermal power plant, such plants are often sited by a river, lake, or estuary which can provide the needed water. This exposes the plant to the risk of flooding. This flooding may conceivably be augmented by the failure of one (or more) upstream dams as a consequence of a major flood. The NRC must therefore be concerned with very extreme floods having return periods far longer than those of concern in most hydrologic and hydraulic work.

Few estimates of flood frequency for return period in excess of 100 years are attempted. The Federal Emergency Management Agency (FEMA) requires estimates of the 500-year flood in connection with its flood insurance program. Builders of dams whose failure would risk substantial loss of life and property usually design such dams with a spillway capable of passing the probable maximum flood (PMF). This flood represents an estimate of the physical upper limit of possible flooding. No attempt has been made to estimate the return period of this flood although one statistical method of estimating the probable maximum precipitation (PMP), from which the PMF is calculated, suggests that the PMP may lie between 5 and 30 standard deviations above the mean daily rainfall [Hershfield, 1965].

Fixing the probability of many kinds of events has become relatively common in today's world. What makes the problem of floods difficult? The problem is basically one of sample size. For analysis of flood probability, we define a flood as the highest instantaneous flow in a 12-month period (October through September) called the water year. The annual flood

BEST COPY AVAILABLE

series, therefore, consists of one flood from each water year of record. Systematic collection of streamflow data began about 100 years ago. A few scattered records of river stage are longer.

The average length of all flow records is less than 50 years, and the number of very long records is quite small. Moreover, the earlier portion of many records are of doubtful accuracy because of poor measuring equipment. Natural and/or artificial changes in the character of many streams and/or their catchments may make many records, including some of the shorter ones, inhomogeneous.

Efforts to determine the appropriate distribution with which to extrapolate from these short records have been inconclusive. The U. S. Water Resources Council [1976] adopted the log Pearson Type III distribution, but the comparisons on which this decision was made did not show any distribution which was clearly superior. The United Kingdom has adopted the extreme value type I distribution for flood frequency analysis [National Environmental Research Council, 1973]. The uncertainties in the analysis of a short historic record preclude any sound conclusions. Indeed, there is no obvious reason why all streams within a political boundary should conform to any one distribution.

The research project described in this report represents an effort to develop a technique which can provide a basis for estimating the flood flows which are associated with probabilities of 10^{-4} or less.

BEST COPY AVAILABLE

II. THE CONCEPT

In 1970 Franz developed a technique for generating hourly rainfall sequences at two or more stations. This procedure permitted the stochastic synthesis of long rainfall records, which could be used as input to a deterministic flow simulation model to produce an equally long record of synthetic streamflow. Ott [1971] used this procedure to generate 500 year records on Dry Creek in California (Fig. 2-1) and on Fisher River in North Carolina. Reasonable results were obtained. The extreme value distribution (Gumbel) appeared to fit the extended record for the California stream best while the log Pearson III seemed to be the best fit in North Carolina. The frequency curves calculated from the synthesized series did not agree with curves calculated from the observed records at the two stations. However, the synthesized flows were within the 95% confidence limits for the observed frequency curve suggesting that the two samples might have come from the same population. Obviously, however, two cases were too few for firm conclusions.

Nasseri [1976] used the same approach to synthesize 400 years of flow at several stations in the French Broad River basin above Asheville, N. C. Again for this eastern U. S. stream the Log Pearson III gave the best fit, with no FIGURE 2-1 significant differences between catchments which varied from 5.5 mi^2 to 945 mi^2. Both studies suggested that 400 to 500 years might still be somewhat too short for the kind of analysis intended.

The studies made a decade ago were costly because of the large amounts of relatively expensive computer time required. Since 1970 there have been new developments in the use of the multivariate normal distribution which was the basis for the multistation stochastic rainfall generator. In addition the computational speed of computers has increased markedly with accompanying reduction in costs. These factors suggest that reconsideration of this approach may lead to some useful

BEST COPY AVAILABLE

BEST COPY AVAILABLE

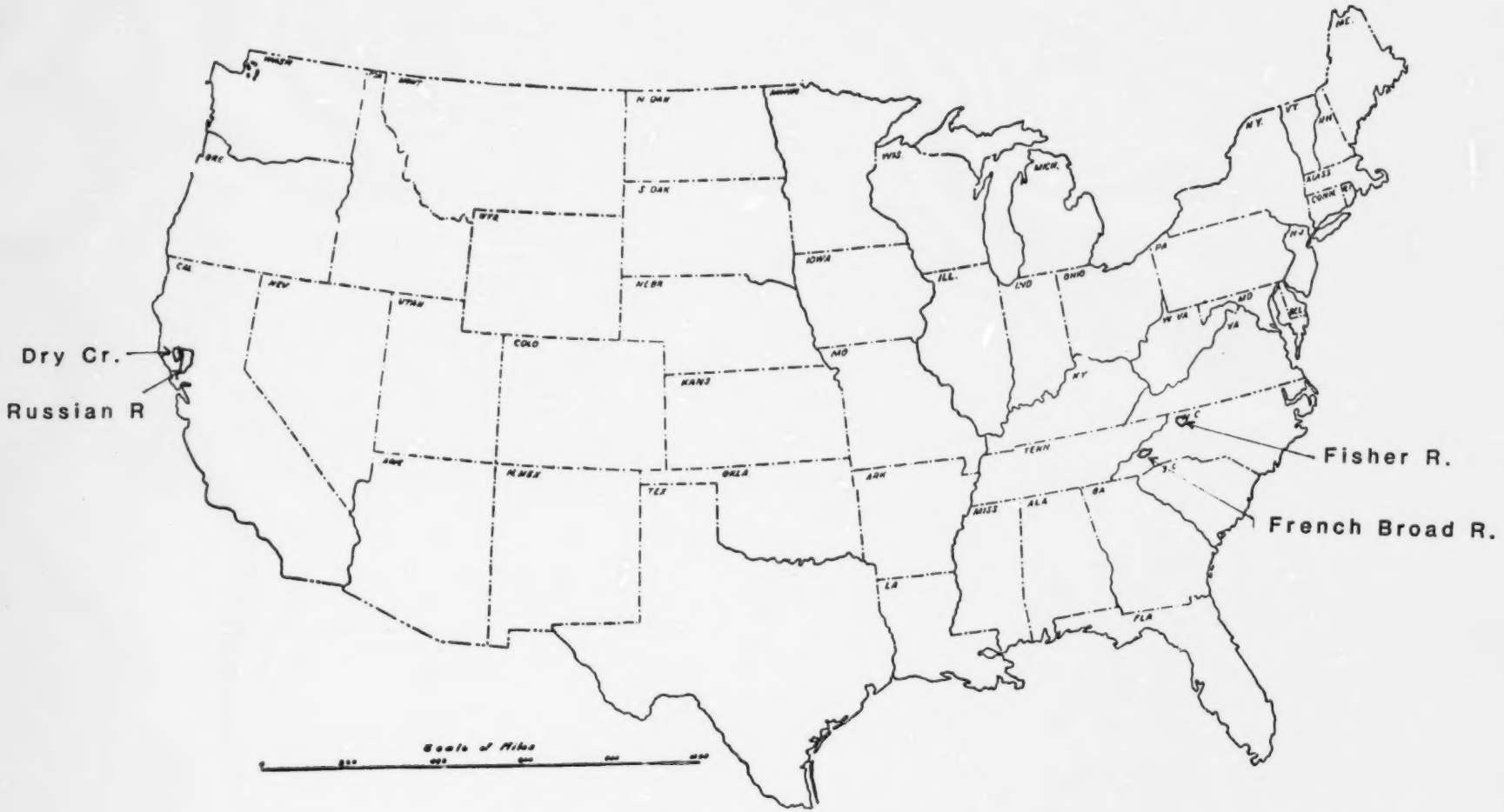

Fig. 2-1 Location of test basins for the concepts

results.

The goal of the preliminary work reported here was to:

- o Develop an improved stochastic rainfall generator using recent developments for the truncated normal distribution.
- o Make use of improved versions of the deterministic flow simulator (SWM)
- o Apply these models to systhesis of 1000 years of flow data on a moderately large catchment.

These steps will permit:

- o A better estimate of the minimum length of systhetic record required for a stable distribution.
- o Test the sensitivity of the results to the changed procedures.
- o Provide still another test of the concept.
- o Provide a basis for planning a more comprehensive test of the concept on a nationwide basis.
- o Permit testing of various methods for evaluating the results for future use.

No basis exists which can unconditionally confirm the reliability of the concept outlined above. There is no way of comparing the synthesized flow series with an historic series of equal length. We can test in the statistical sense, i.e., show that the synthesized flows come from the same population as the historic flows. It can, of course, be shown that the rainfall generator reproduces the parameters of the historic record and that the streamflow simulator can reproduce the historic flows within reasonable tolerances. In the end the reliability of the methodology and the logic of the concept must be convincing.

BEST COPY AVAILABLE

III. THE HYDROLOGIC SIMULATION MODEL

For the simulation of streamflow a version of the Stanford Watershed Model (SWM) [Crawford and Linsley, 1966] has been used. This version is called SWM84 [LKAssociates, 1985] programmed by Linsley, Kraeger Associates for use on their Cromemco computers. Some modifications of the original model have been made to include new concepts in hydrology and to provide more useful output to users. A flow diagram of SWM84 appears as Fig. 3-1.

Hourly rainfall data and monthly potential evapo-transpiration are input to the model along with values for 18 parameters which describe the geometry and physical characteristics of the catchment. Each hour of rainy days is processed through the algorithms of SWM84 to yield hourly values of runoff to the stream. Rainless days are treated as aggregates of 24 hours. The runoff values are the sum of runoff from impervious surfaces, overland runoff from pervious areas, interflow, and groundwater flow to the stream.

The computed increments of runoff are routed through the channel system of the catchment using kinematic routing methods, to determine the hydrograph of outflow. The system can run from any set of initial conditions for as long as rainfall and potential evaporation are input.

The parameters defining the basin geometry can be determined from maps or aerial photographs, but four parameters defining hydrologic characteristics of the catchment are best found by calibration using observed rainfall and streamflow data.

BEST COPY AVAILABLE

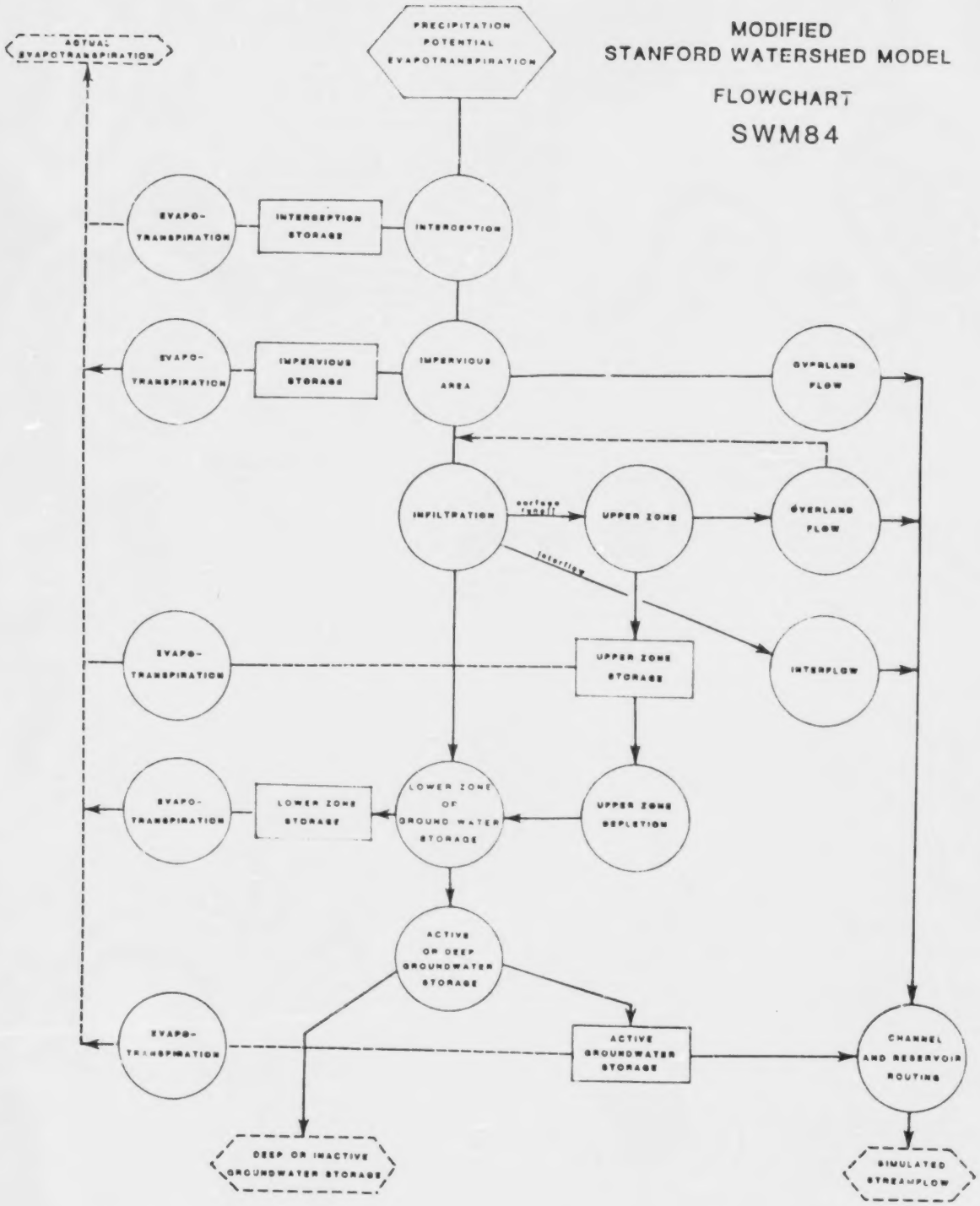

Fig. 3-1 Flow diagram for SWM84

BEST COPY AVAILABLE

IV. THE STUDY CATCHMENT

The Russian River near Cloverdale, CA was chosen as the study basin (Fig. 4-1). The Russian R. enters the Pacific Ocean 65 miles north of the Golden Gate, after flowing south for most of its course in a valley east of the Coast Range Mountains (Fig.4-2). The characteristics of the catchment and of the tributary catchments and the basins upstream of Cloverdale are given in Table 4-1.

Table 4-1. Characteristics of the Study Basin

Stream	Station	Drainage Area sq.mi.	Mean Rain inches	Annual Runoff inches	Record
Russian R.	Cloverdale	503	42.2	22.0	1951-
Russian R.	Hopland	362	42.4	20.7	1939-
Russian R.	Ukiah	99.7	42.8	24.1	1952-
E.Fk.Russian R.	Calpella	93.0	39.5	24.8	1941-
E.Fk.Russian R.	Ukiah	105	39.5	23.3	1951-
Feliz Cr.	Hopland	31.1	46.0	20.3	1958-66

The catchment has two features which complicate the analysis. Some 300 cfs are diverted from the Eel River basin to a powerhouse at Potter Valley in the upper East Fork catchment. This water is then subject to diversion for irrigation, and only the return flow from irrigation and the water not used for irrigation continues in the Russian River. Lake Mendocino on the East Fork is a flood control reservoir built by the U. S. Corps of Engineers. By regulating outflows during floods and releasing water for flow maintenance during the summer, this reservoir alters the natural streamflow below it. The evaporation from the reservoir also reduces total flow downstream from it.

The stream gage on the Russian R. at Ukiah permitted the catchment above Cloverdale to be simulated in two portions. The rain gage at Redwood Valley was used for the northern segment and parameters for this segment were determined by

BEST COPY AVAILABLE

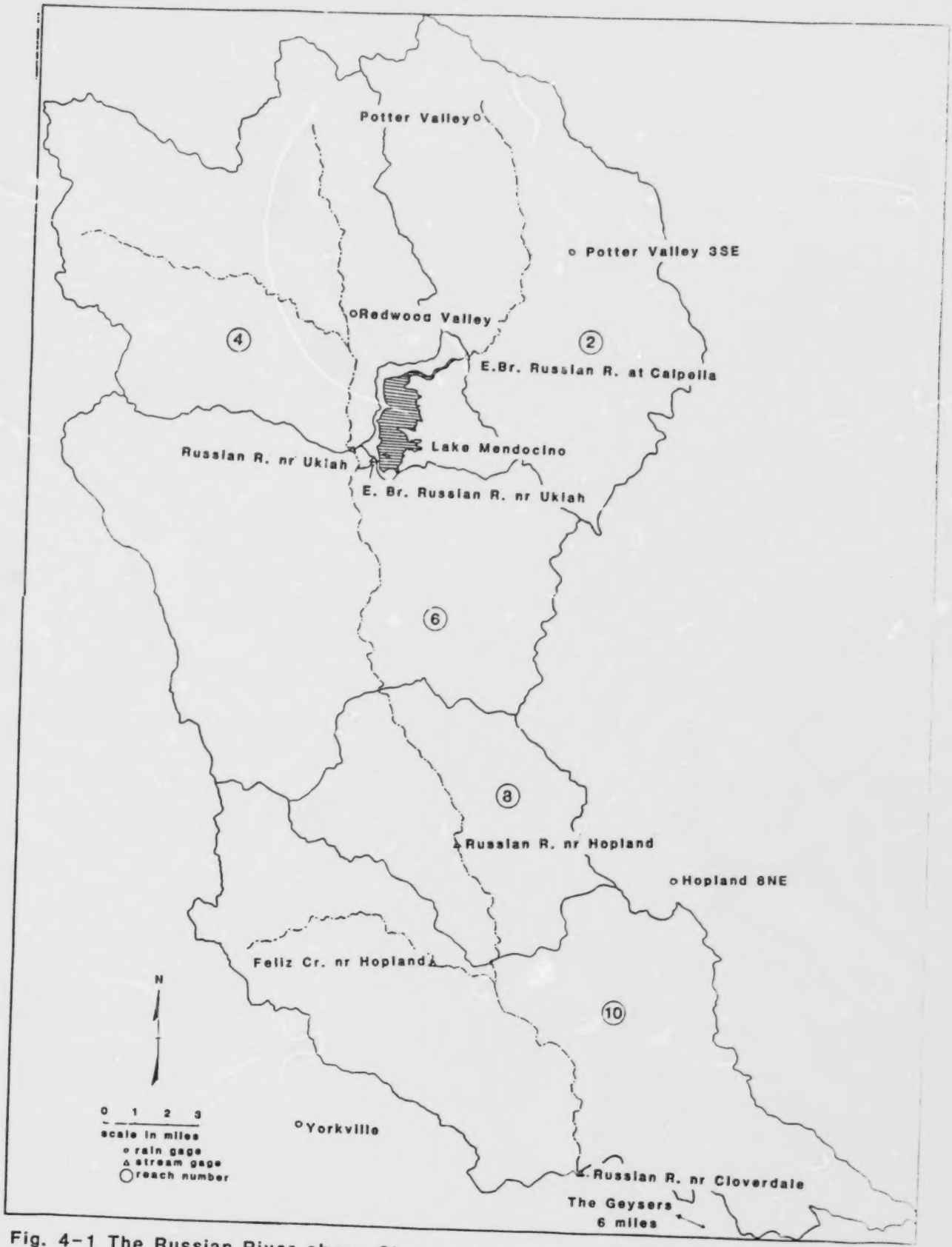

Fig. 4-1 The Russian River above Cloverdale showing reaches and station locations

BEST COPY AVAILABLE

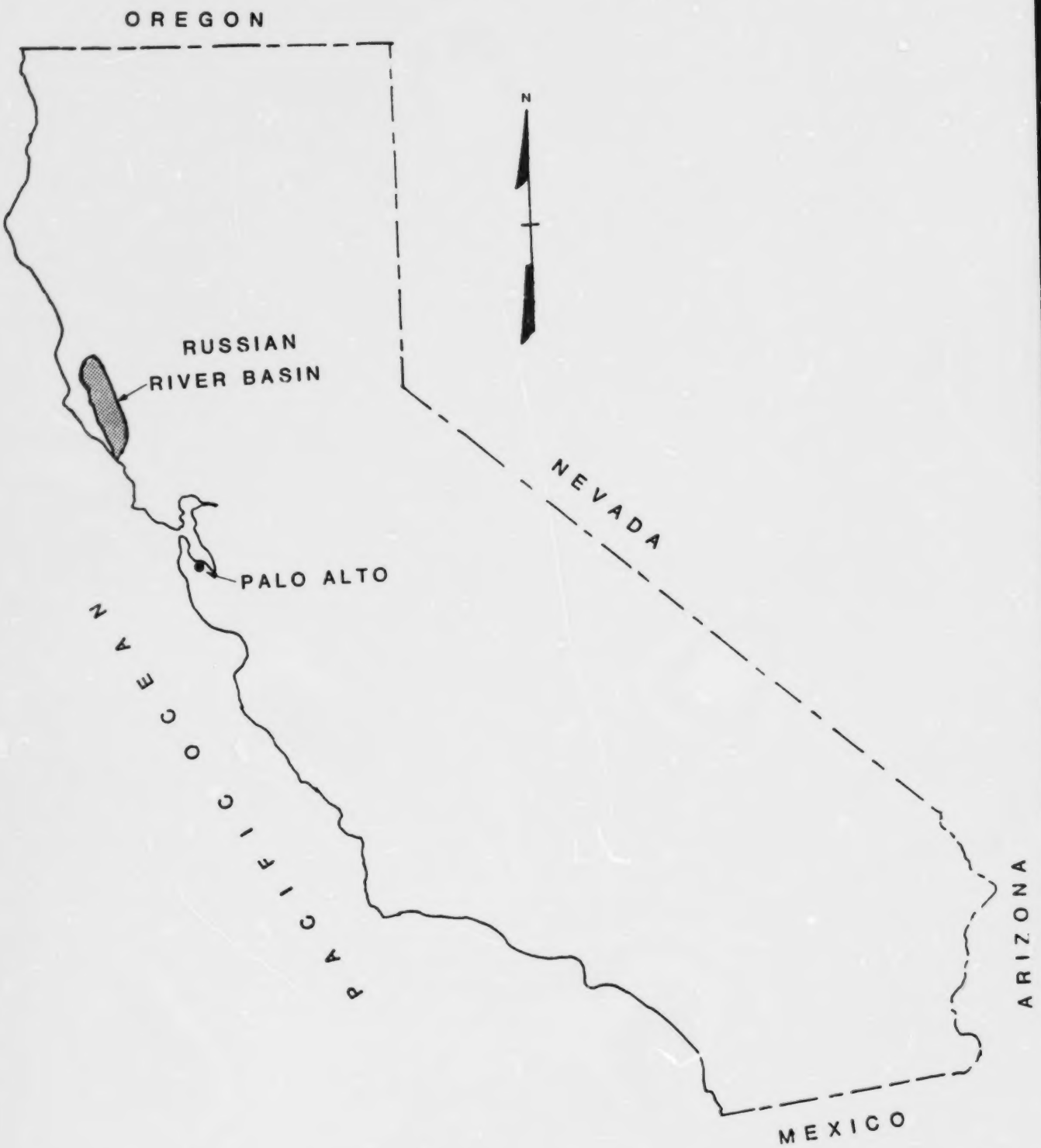

Fig. 4-2

Map showing the location of the study basin in California.

Best COPY AVAILABLE

calibration against the gage at Ukiah. Unfortunately the Redwood Valley station was discontinued in 1972, thus limiting the calibration period to the years 1952 through 1972. Yorkville rainfall was used to represent the southern segment.

Table 4-2 lists the parameters used to synthesize the flows for the Russian R. nr. Cloverdale for the calibration period, October 1952 through September 1972.

Figure 4-3 shows the comparison of the observed and synthesized flood frequency curves for the station near Cloverdale. This is probably the most important test since it confirms that the model closely reproduces the historic flood probability.

Figures 4-4 through 4-6 compare simulated and observed hydrographs of mean daily flow at Cloverdale. Only a few relatively wet years are included in this report. These plots show that all major peaks are simulated closely and that low flows are represented well. The agreement for the

Table 4-2

Parameters for the Russian River above Cloverdale

		North	South
K1	Rainfall Adjustment	1.40	0.81
A	Impervious fraction	0.15	0.01
EXPM	Interception storage pervious	0.10	0.10
UZSN	Upper zone storage	0.80	0.80
LZSN	Lower zone storage	10.00	10.00
K3	Evaporation parameter	0.50	0.50
K24L	Seepage to groundwater	0.00	0.00
K24EL	Evaporation from groundwater	0.00	0.00
INFILT	Infiltration	0.10	0.05
INTFLW	Interflow	2.50	2.50
LP	Length ofoverland flow - pervious	500.00	500.00
SP	Slope of overland flow - pervious	0.10	0.10
NNP	Manning's "n" - pervious	0.35	0.35
IRC	Interflow recession constant	0.35	0.35
KV	Variable groundwater recession	1.00	1.00
KK24	Groundwater recession constant	0.98	0.98
LI	Length overland flow - impervious	100.00	100.00
SI	Slope overland flow - impervious	0.01	0.01
NNI	Manning's "n" -impervious plane	0.03	0.03
EXPMI	Interception - impervious plane	0.05	0.05

lesser peaks is less close, mainly because rainfall reported at two stations during small storms is often not representative of the average rainfall over the relatively

BEST COPY AVAILABLE

BEST COPY AVAILABLE

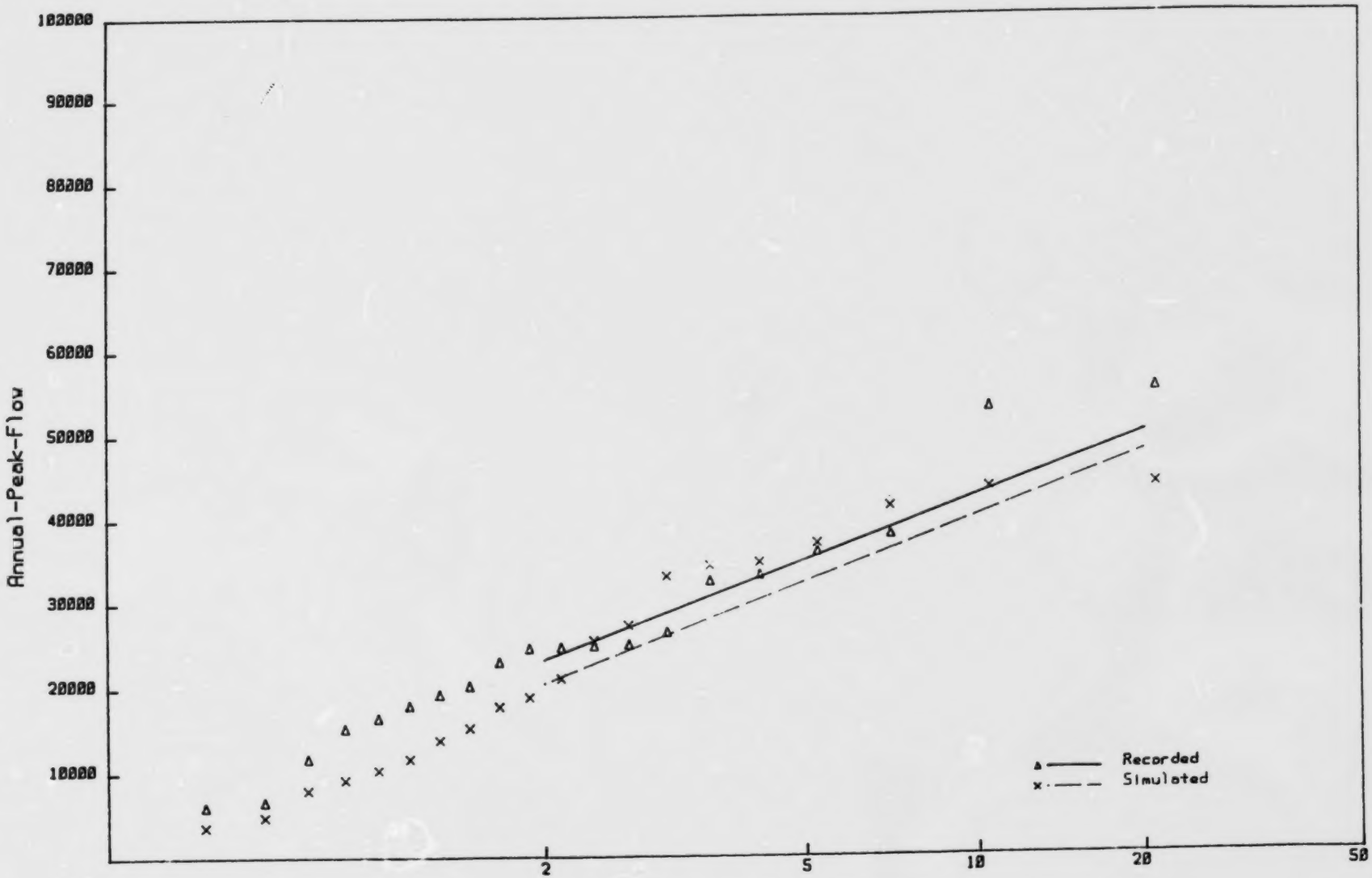

Fig. 4-3 Comparison of observed and simulated flood probabilities,

Russian River nr Cloverdale, CA., 1952-72

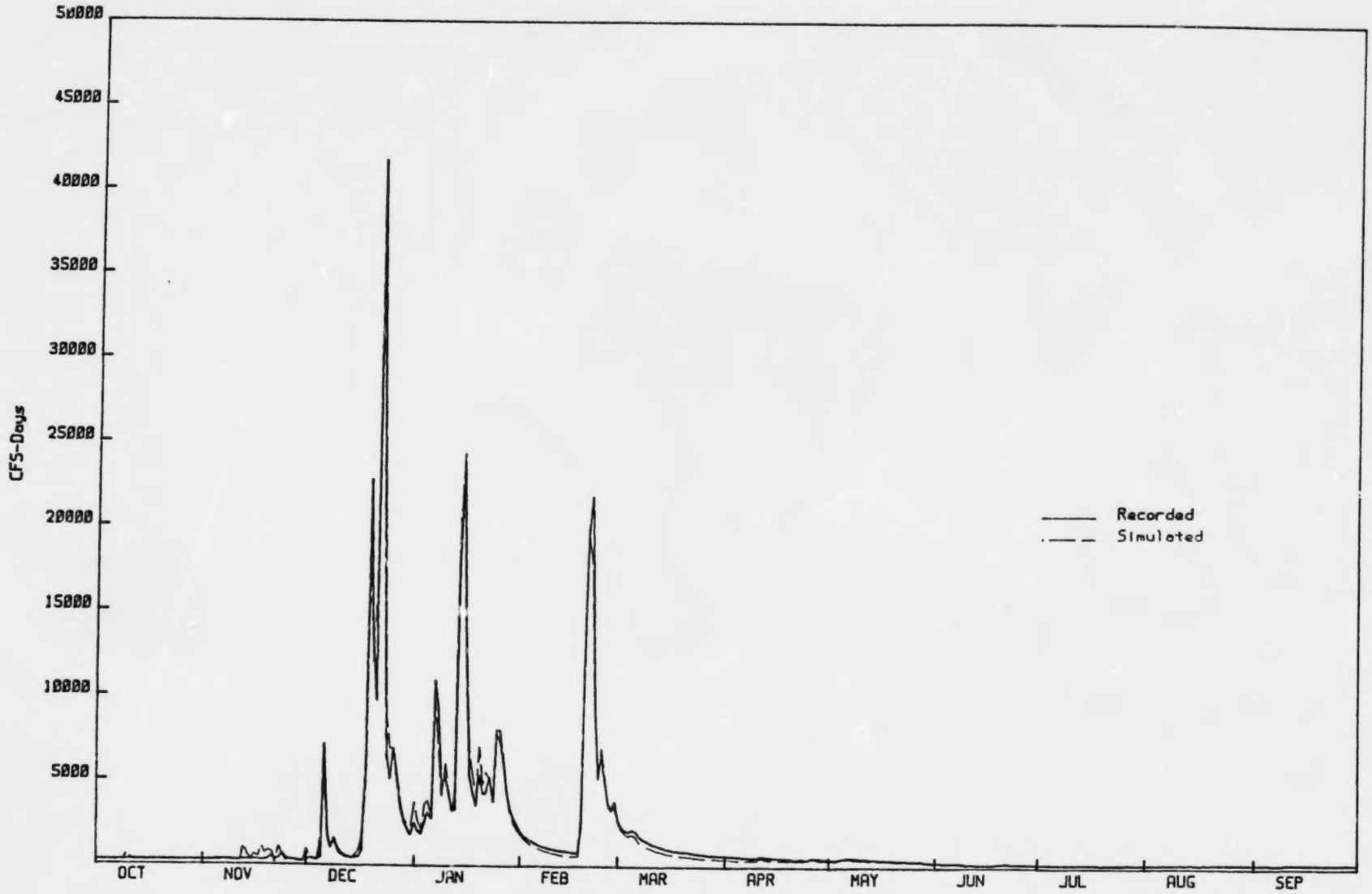

Fig. 4-4 Plot of simulated and observed mean daily flows at Cloverdale for the water year 1956

BEST COPY AVAILABLE

BEST COPY AVAILABLE

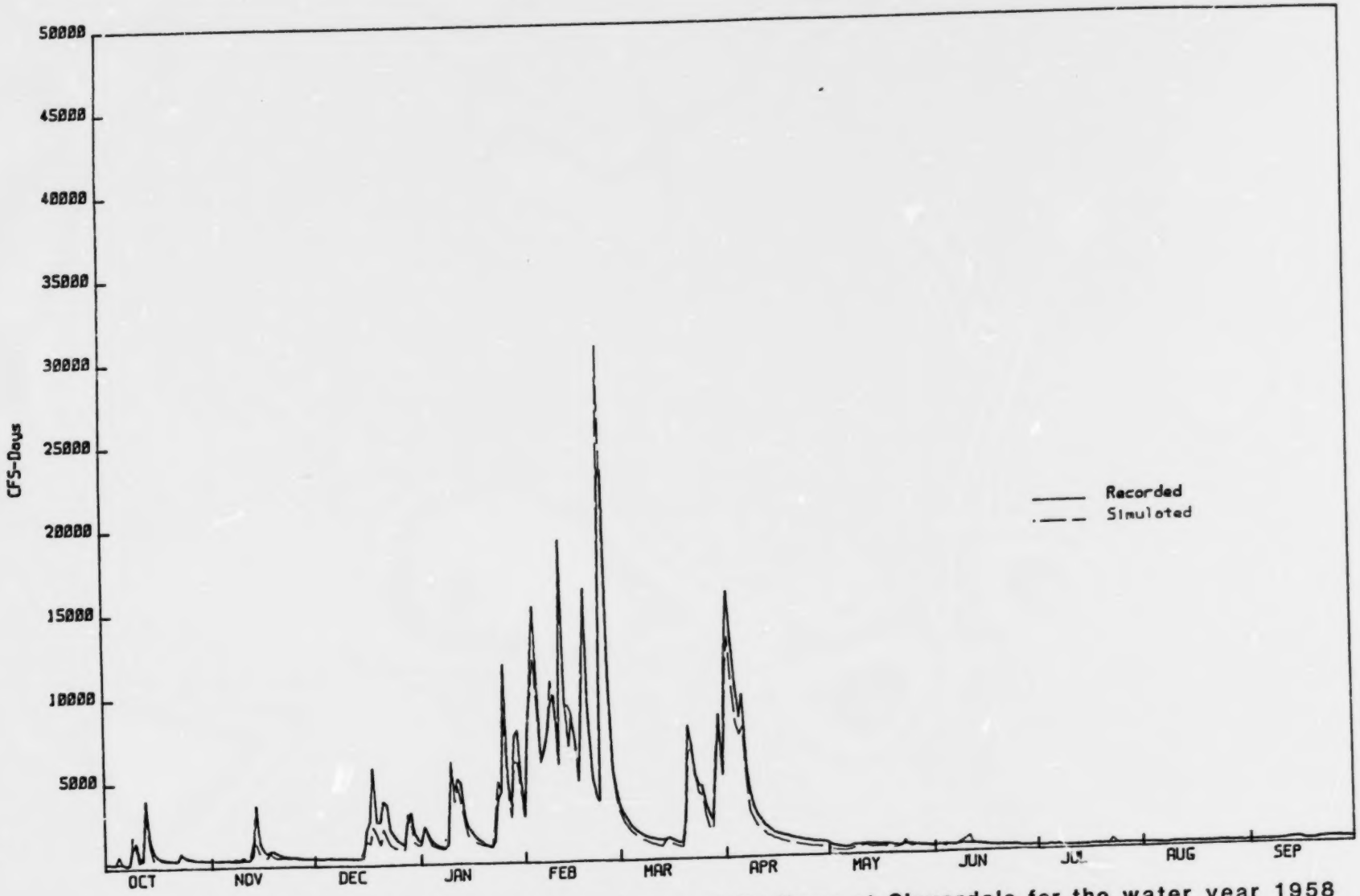

Fig. 4-5 Plot of simulated and observed mean daily flows at Cloverdale for the water year 1958

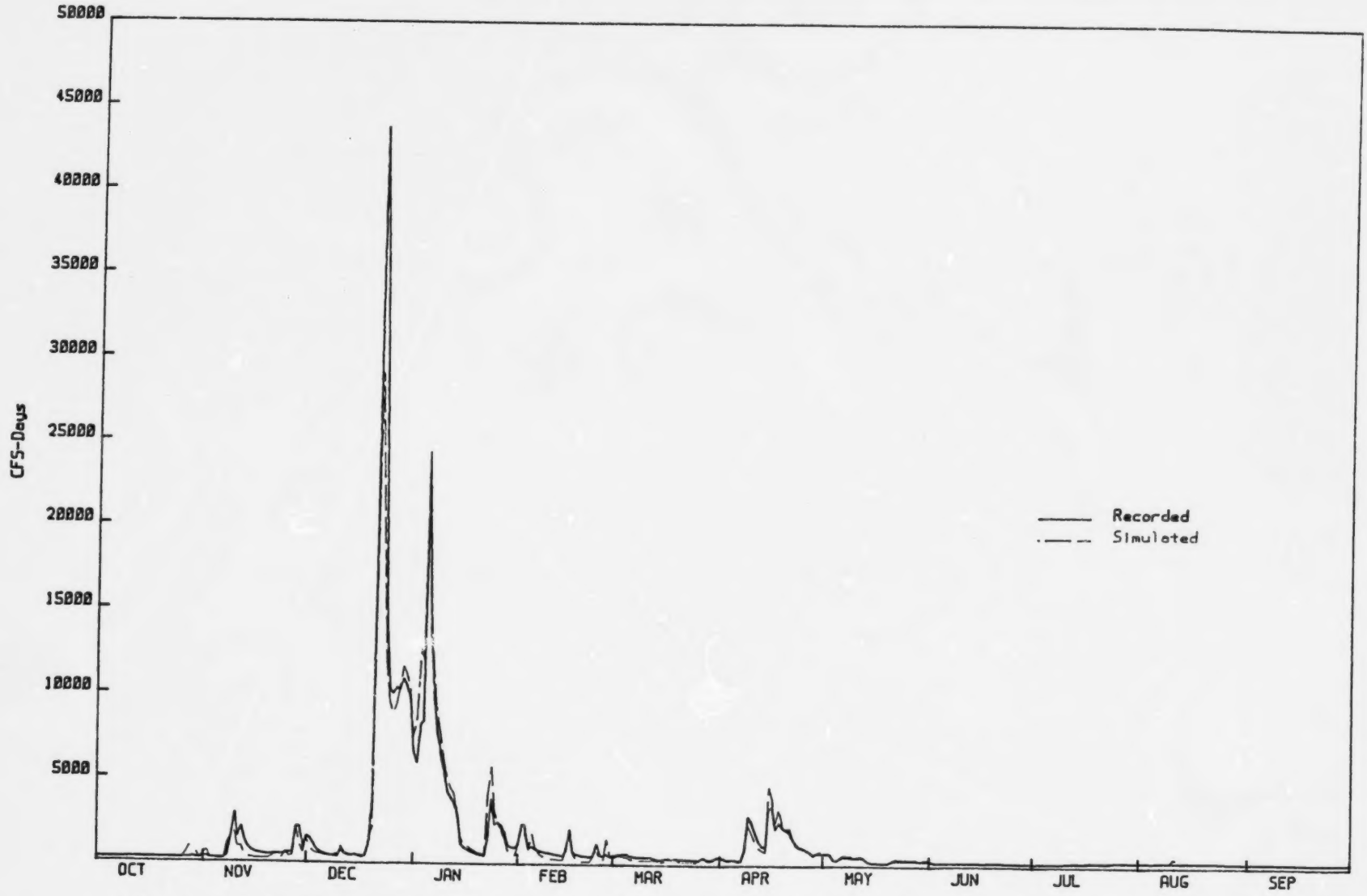

Fig. 4-6 Plot of simulated and observed mean daily flows at Cloverdale for water year 1965

BEST COPY AVAILABLE

large catchment.

Figures 4-7 through 4-10 contain plots of the hourly streamflows for selected major storms. The fit of these more detailed hydrographs is not as good as the plots of mean daily flows seem to indicate. For purposes of calibration, outflows from Lake Mendocino as measured by the gage on the East Fork near Ukiah have been input in lieu of simulating these outflows. This step eliminates errors that might be introduced by using assumed operation and accurately reflects the effect of the lake on flood flows at Cloverdale with one exception. However, when flow occurs over the ungated spillway, it returns to the river downstream of the gage on the East Fork Russian River near Ukiah and is unmeasured. This is believed to explain much of the difference between observed and simulated flows near Cloverdale for the 1964 flood. In addition the use of only two raingages to simulate the 503 mi^2 catchment does not seem to adequately represent the rainfall fluctuations over the catchment. Nevertheless, for the 20-yr calibration period the historic flood frequency is closely matched.

Table 4-3 presents a comparison of observed and simulated annual flow volumes for the station near Cloverdale. The annual volume of runoff is quite accurately simulated in most years. The maximum percentage error in annual runoff is 17% (3.38 in.). The maximum absolute error is 4.62 in. (10.24%). Over the 20-yr period the mean is simulated 0.75 inches low (2.9%) and the average percentage error is 5.8%. Although the study is aimed at flood peaks, a reliable estimate of peaks is not possible if there are gross errors in the simulation of flow volume. The flow duration curve (Fig. 4-11) also confirms that the total volume of runoff and the distribution of flow rates are closely simulated.

BEST COPY AVAILABLE

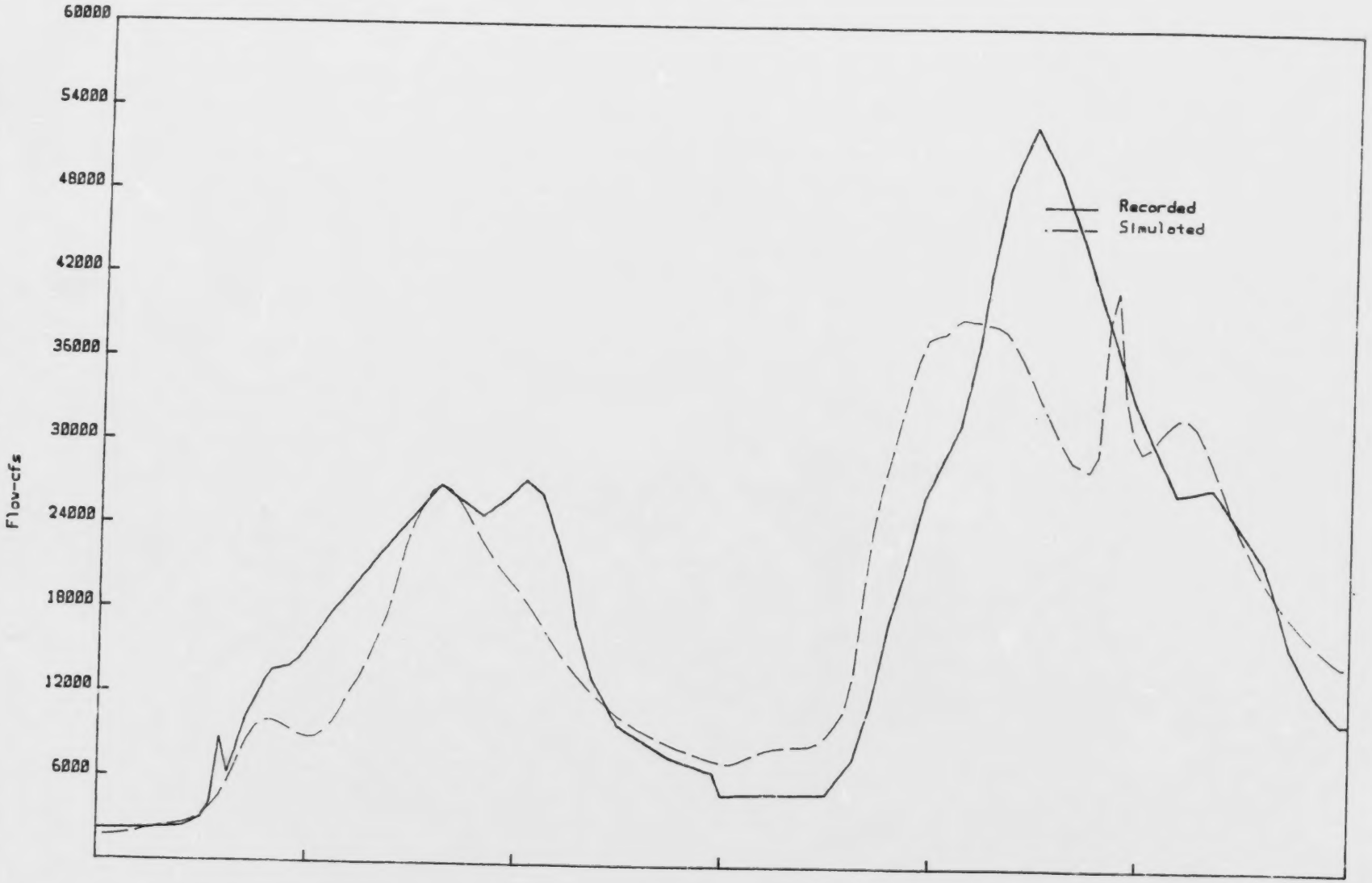

Fig. 4-7 Observed and simulated hydrographs for the flood December, 1955 at Cloverdale

BEST COPY AVAILABLE

BEST COPY AVAILABLE

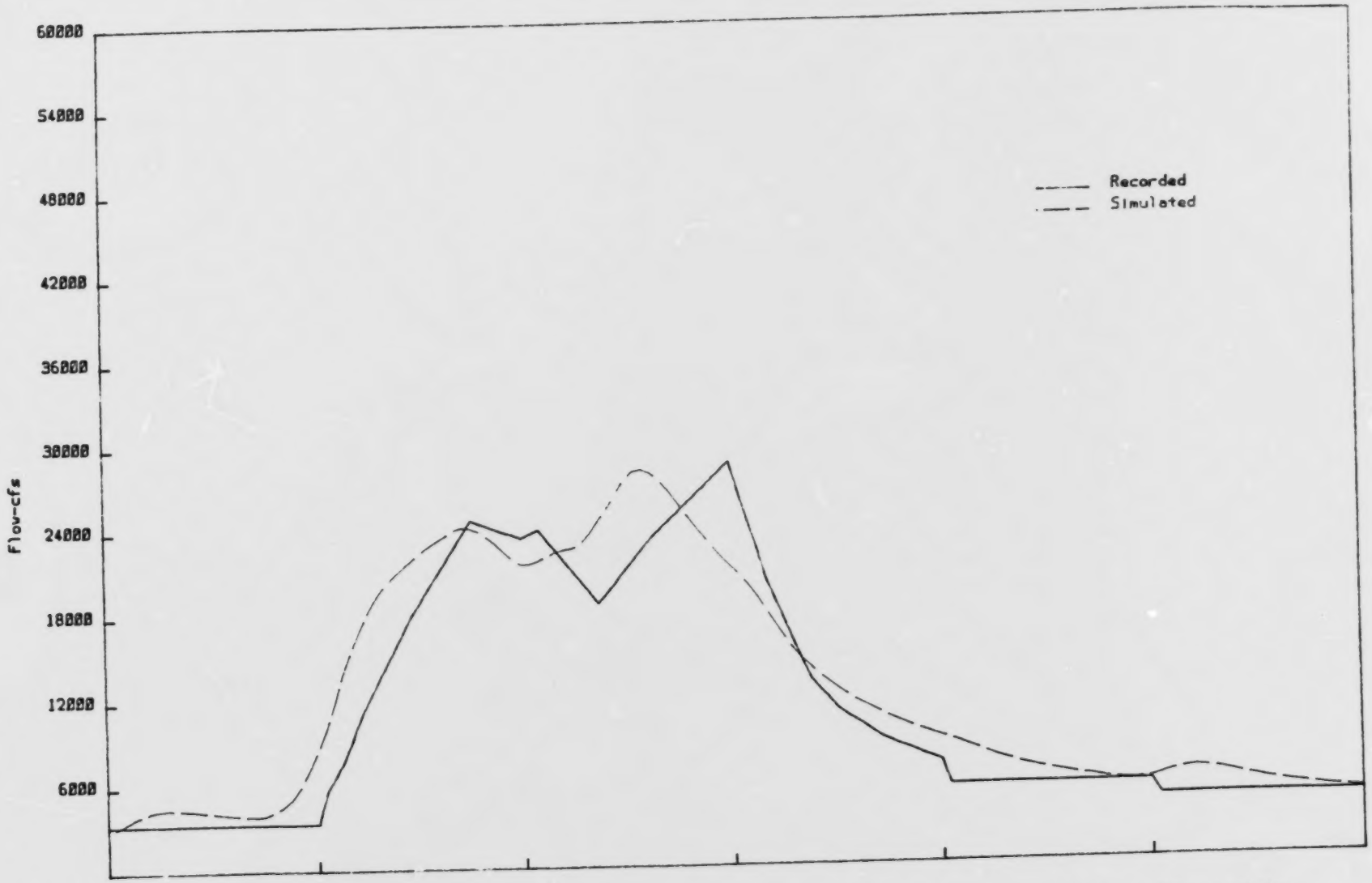

Fig. 4-8 Observed and simulated hydrographs for the flood of January, 1956 at Cloverdale

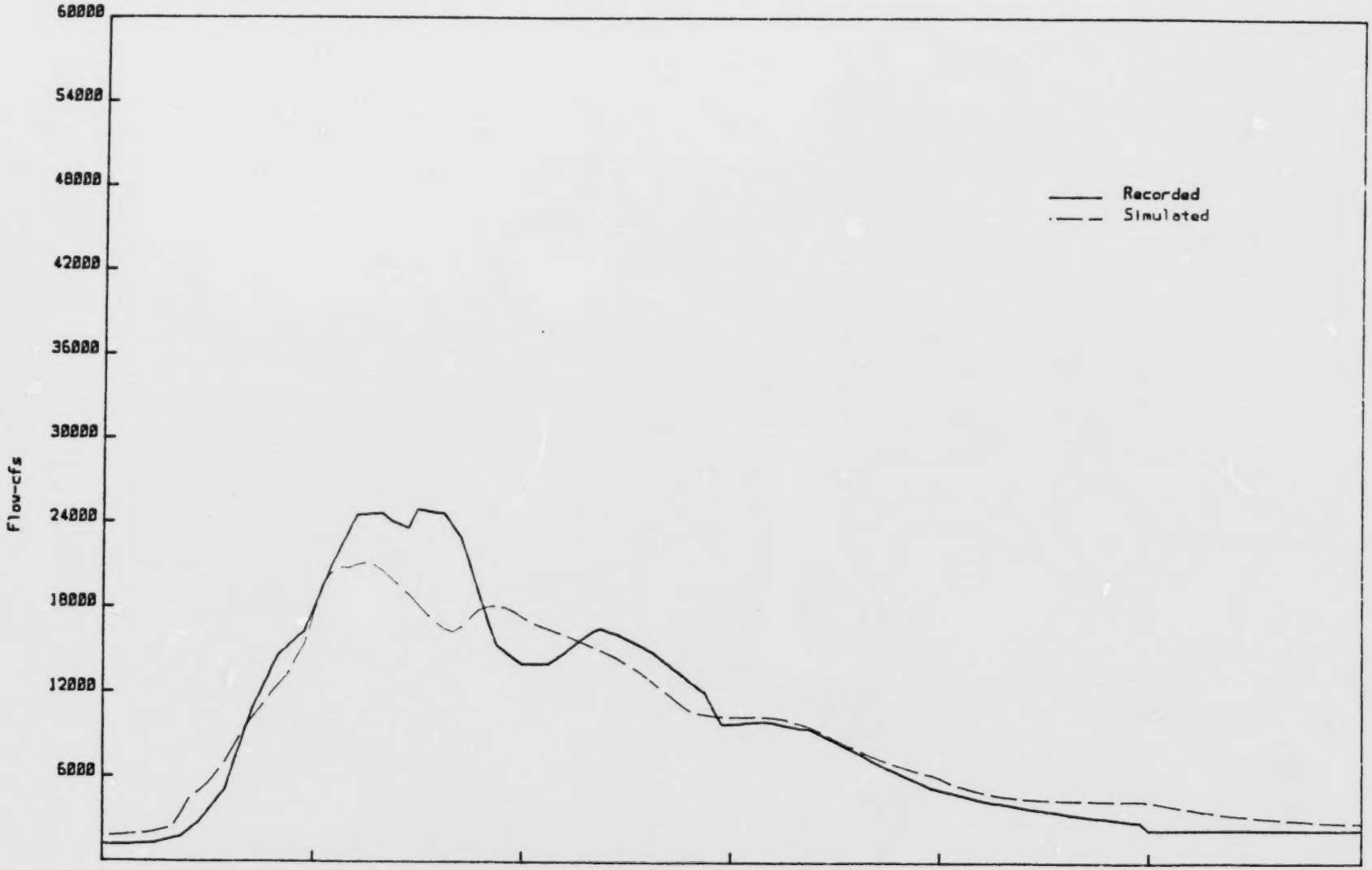

Fig. 4-9 Observed and simulated hydrographs for the flood of Febuary, 1960 at Cloverdale

BEST COPY AVAILABLE

BEST COPY AVAILABLE

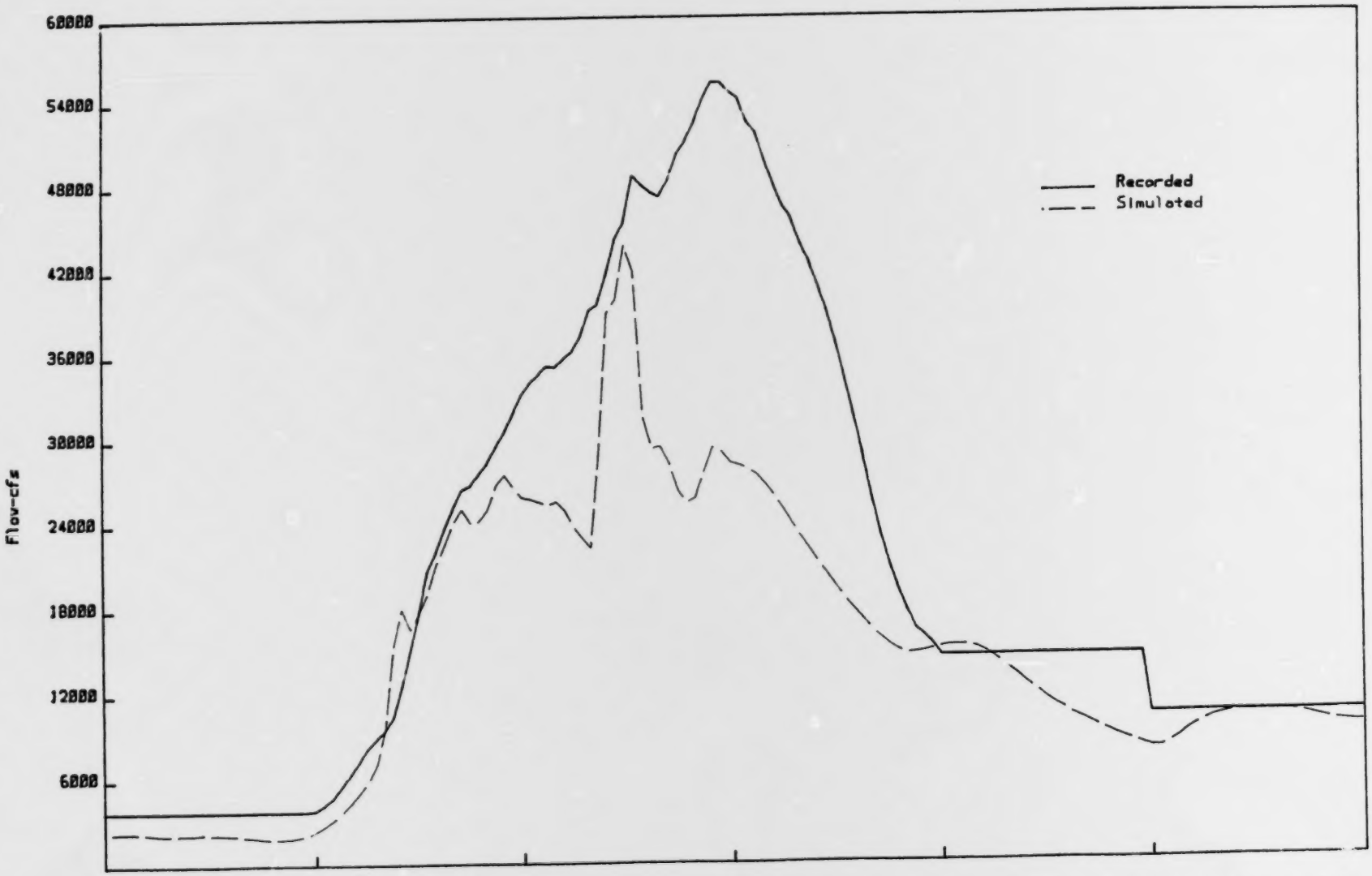

Fig. 4-10 Observed and simulated hydrographs for the flood of December, 1964 at Cloverdale

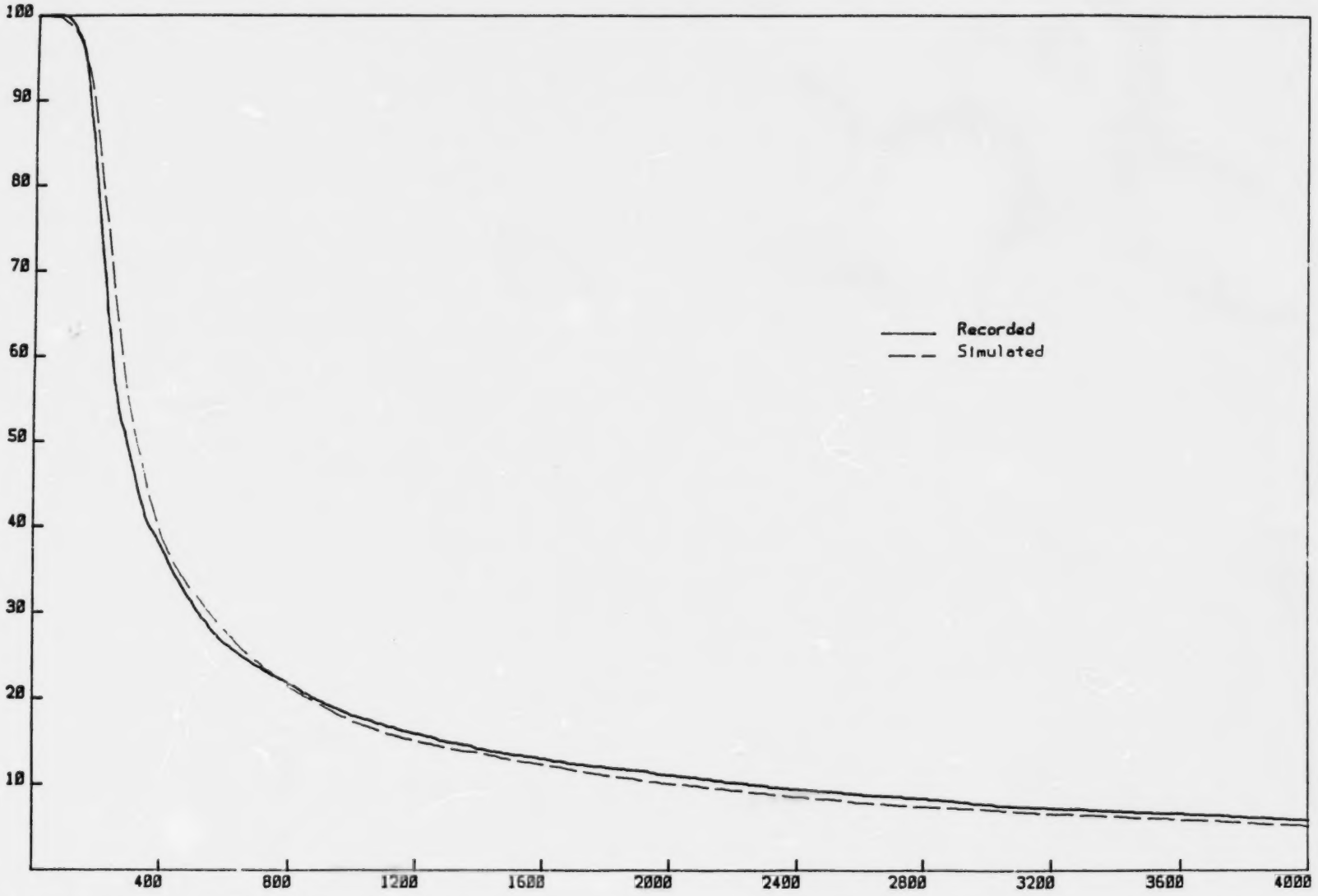

Fig. 4-11 Observed and simulated flow duration curves for the Russian River near Cloverdale, 1952-72

BEST COPY AVAILABLE

Table 4-3

Observed vs. Simulated Annual Runoff

Russian River, near Cloverdale, CA.

year	Observed (inches)	Simulated (inches)	Difference (%)
1952-53	29.81	29.90	0.29
1953-54	25.20	22.69	-10.24
1954-55	14.28	13.45	-5.87
1955-56	41.77	42.38	1.47
1956-57	18.54	17.98	-3.02
1957-58	50.15	45.97	-8.33
1958-59	14.64	14.88	1.63
1959-60	18.68	20.67	10.65
1960-61	22.15	23.03	3.94
1961-62	20.75	21.67	4.43
1962-63	28.00	25.61	-8.52
1963-64	14.14	13.33	-5.73
1964-65	35.50	34.22	-3.61
1965-66	23.48	23.16	-1.39
1966-67	31.25	27.92	-17.06
1967-68	18.33	15.43	-15.80
1968-69	38.70	39.85	2.97
1969-70	34.52	35.72	3.49
1970-71	27.91	26.64	-4.54
1971-72	13.15	13.42	2.03
Average	26.05	25.30	-2.9%

BEST COPY AVAILABLE

V. THE STOCHASTIC RAINFALL

The goal in generating stochastic hourly rainfall data was to mimic the major chacacteristics of observed hourly rainfall in the study catchment. The characteristics of interest include:

- o probability of a storm
- o probability distribution of the length of storm periods
- o probability distribution of the hourly rainfall amounts
- o coincidence of rainfall at adjacent stations
- o persistence of the hourly rainfall amounts

Other characteristics such as the monthly and annual totals are also of interest.

The fundamental approach taken is elementary in that the model makes no explicit use of the physical processes involved. Therefore the model is purely stochastic. The model is designed to produce a series of numbers which "look" like hourly rainfall data. What hourly rainfall data looks like must be defined by analysis of observed records. Various tests can be applied to establish suitable agreement for a finite number of comparisons. Obviously, the generated numbers are not and cannot be rainfall data. We are, however, interested in agreement that is sufficiently close for only a subset of the nearly infinite number of tests that can be devised.

This approach is similar to that used in generating so-called uniform random numbers on a digital computer. These numbers are not random at all and are in fact exactly reproducible. Sequences of these numbers have many, but not all the characteristics of a random series. An exact operationally meaningful definition of what constitutes a random series of numbers is difficult and Knuth [1981] has a highly recommended discussion of this topic. The goal in generating these psuedo-random uniform numbers on the computer is to mimic the important characteristics that we want a

BEST COPY AVAILABLE

random series to have. A wide variety of tests have been devised to test for these characteristics and Knuth is a good source for the interested reader.

Because the model is purely stochastic, it becomes important to have as many explicit relationships between the characteristics of the observed hourly rainfall and the parameters of the model as possible. This is the only way to ensure that these characteristics will be properly mimicked. As an example, the model developed by Franz (1970), had no explicit relationship between the distribution of storm lengths and the parameters. It happened that the storm length distribution was reproduced well in four test cases and it appears that the storm length distribution was implicit in the parameters. However, the evidence is sparse and it is far more comforting to be able to estimate parameters which will always produce the desired result. In general only a few of the characteristics of the observed hourly rainfall can be mimicked explicitly by the stochastic model. The other characteristics are then used to verify the adequacy of the model.

Related Work

Three types of stochastic models for rainfall have been proposed.

1. Models which represent selected seasons of the year or selected storm types (Duckstein et [1972] and Eagleson and Grace [1966]).

2. Models which represent rainfall at all points and times over a limited region and short time span (Bras and Rodriguez [1975]).

3. Models which generate short-interval rainfall at one or more specific locations for a long time period.

It is the third class of models with which this project is concerned. This class of models produces synthetic

BEST COPY AVAILABLE

rainfall sequences at the location of existing raingages. All storms and the intervals between storms are produced by these models. A hydrologic model must be calibrated to a catchment using concurrent records of rainfall and streamflow. The location and nature of the recorded rainfall influences the parameter set developed for the model. It is, therefore, important to develop synthetic rainfall sequences which mimic the characteristics of the rainfall recorded at a specific site.

The first model of class three was developed by Pattison [1965]. He used a Markov chain to represent the hourly rainfall at a single station. He found that the state of a given hour (wet or dry) had to be determined by the states of the previous six hours. Once the hour was known to be wet then only the previous hour's rainfall was necessary to define the amount of rainfall. Using the synthetic rainfall as input to the Stanford Watershed Model, he was able to show that within the sampling uncertainty of short records, that the flood peaks from the synthetic rainfall were from the same population as the flood peaks observed for the stream.

Extension of the Pattison approach to more than one station is impracticable because the number of required parameters becomes large. Also possible transitions which were not observed in the period of record cannot be in the synthetic record and new extremes cannot be synthesized.

Franz [1970] introduced a model which overcame most of the problems with the Pattison approach. He saw that any practical model would have to be based on the multivariate normal distribution. No other convenient multivariate distribution exists. Such a model was not unduly complicated if the persistence structure of the hourly rainfall was assumed to be Markovian.

Hourly rainfall is <u>not</u> normally distributed. It is never negative and most of the time it is zero. Normal autoregressive models do not produce sequences which look at all like observed data. A power function transformation of the hourly rainfall was developed to make its distribution

BEST COPY AVAILABLE

more nearly normal. The many zeros were accomodated by developing submodels for storm periods and for interstorm periods. A storm period was defined as any sequence of consecutive hours in which at least one station had non-zero rainfall. Interstorms were the periods during which all stations had zero rainfall.

Hourly rainfall is an intermittent process and there were still zeros in the storms periods defined as above. An intermittent process was created from the continuous process by viewing the observed sequence of hourly rainfall as being the result of filtering the negative values from an unobservable background process. The filter replaced every negative value with a zero value to produce the hourly values. The parameter estimation process involved concepts from probabililty plotting and least squares and was somewhat cumbersome.

Results with this model were generally good. The division of the year into seasons to account for the annual variability in rainfall introduced some discrepancies. Some adjustments to the rainfall amounts were needed to correctly reproduce the 0.01 inch/hour probabilities. The distribution of storm lengths and annual and monthly rainfalls were replicated well. None of these values were explicitly used in defining the parameters of the model.

Ott [1972] used the Franz model to study flood frequency distributions on Dry Cr. in northern California and Fisher River in North Carolina. These represented two very different rainfall regimes. He found that the treatment of zero hours at the beginning and end of a storm had to differ between the two areas. He used a weighted least squares fit for the normalizing transformation on Dry Cr. and an unweighted fit on the Fisher R.

Nasseri [1976] applied the same approach to the French Broad River in North Carolina. He looked at 10 subwatersheds above Asheville, N.C. to see if there was any effect of catchment size and found none.

BEST COPY AVAILABLE

Kelman [1977] used the same assumptions and structure as the storm submodel of Franz to synthesize daily rainfall. He found that he could dispense with a separate interstorm model. Kelman estimated parameters with a combination of maximum likelihood and moment estimators. The zero days were included in the estimation process. His approach is preferable to Franz' solution because the moments of the observed rainfall enter the estimation process directly. His process also permits the simultaneous estimation of the parameters of a given site (mean, variance, transformation constant, and serial correlation). Kelman's tests show that his model did well on the rainfall which would contrbute to floods, but the distribution of runs of consecutive days of rain was distorted for the shorter runs.

Richardson [1977] applied the Kelman model to a 20-station network in northeastern Texas. He concluded that "....the truncated multivariate normal distribution is a useful model of daily precipitation over an area if the precipitation at a point can be transformed to approximate a truncated normal distribution." Richardson used the square root of rainfall as the normalizing transformation. A Fourier series with six harmonics was used to represent the annual variations of the means and standard deviations. He found evidence that the lag-one Markov dependence structure did not describe the autocorrelations for lags greater than one day.

Description of the Model

The characterstic of hourly rainfall which dominates the modelling process is that it is not raining most of the time. Thus hourly rainfall data are intermittent. There are basically long periods with no rainfall punctuated with relatively short bursts of rainfall. The wettest season of the year in the study network has rain about one hour out of five on the average. A mid-summer season only recorded two hours with rainfall in a 17-year period of record. This shows that another characteristic of rainfall is that it varies greatly throughout the year. Both the intermittent character

BEST COPY AVAILABLE

and seasonal variation must be included explicitly in the stochastic model.

The seasonal variation was incorporated by subdividing the year into shorter periods called seasons. The goal in selecting the seasons is to have the season short enough so that the rainfall characteristics are constant but yet long enough so that there are sufficient observations for determining what these characteristics are. In the current case a mean water year (beginning October 1) of length 8766 hours (leap year: 8784 hours, non-leap year: 8760 hours) was divided into 18 periods of 487 hours each. Five of these periods were aggregated in order to obtain sufficient data for defining the rainfall characteristics during the dry season. Thus 14 seasons were used, 13 of length 487 hours and 1 of length 2435 hours.

The intermittency of the rainfall process was included by hypothesizing that the non-zero observations of hourly rainfall resulted from the filtering of a non-observable background process. This background process produces a value for each hour but rain occurs only when the value for a given hour is greater than some threshold. All values at or below the threshold become zero and represent the dry periods between rainfalls. All values above the threshold are transformed such that they look like the hourly rainfall values (Fig. 5-1). This approach was used by Franz [1970], Kelman [1977], and Richardson [1977]. It appears to be one of the few workable methods for incorporating intermittency explicitly in the stochastic model.

Single Station Model

The model at a single station is represented as follows. For each hour,i, let

$$X_i = D_1 X_{i-1} + \ldots + D_n X_{i-n} + D_0 Z_i \qquad (5\text{-}1)$$

If $x_i > h$ then

$$R_i = A_1 (X_i - h)^p + A_2 (Y - h)^q + 0.5 \qquad (5\text{-}2)$$

else

BEST COPY AVAILABLE

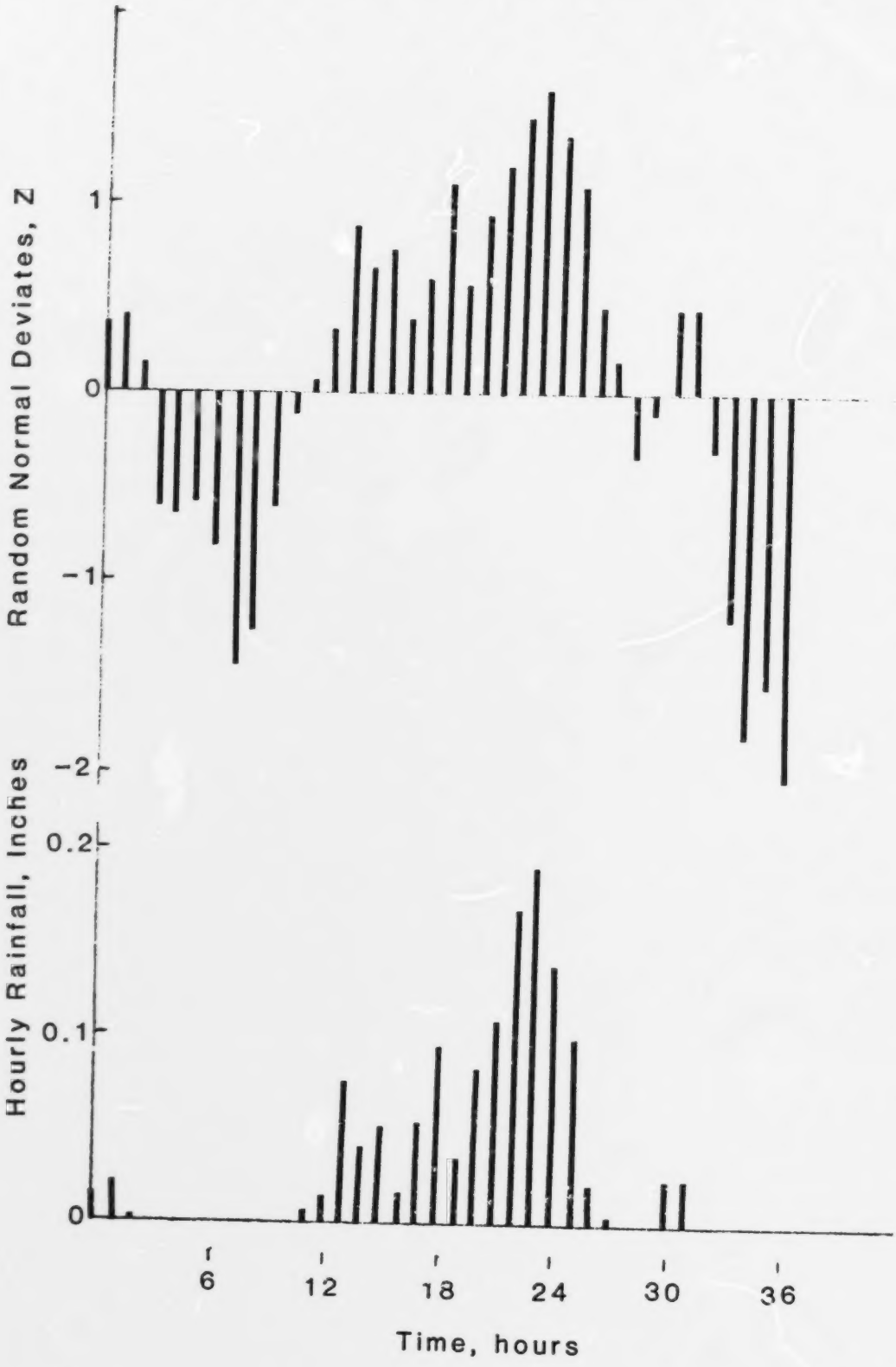

Fig. 5-1 Example of Rainfall Generating Scheme

BEST COPY AVAILABLE

$$R_i = 0 \qquad (5\text{-}3)$$

In the above equations X_i= background process value for the i-th hour, R_i = rainfall value in the i-th hour, Z_i = standard normal random variable, Y = truncated standard normal random variable with $Y > h$, $D_0 \ldots, D_n$ = parameters of the background process, h = threshold or truncation level for the background process, A_1, A_2, p, q = transformation parameters for defining rainfall amounts. The random variables, Z_i and Y, are distributed independently of X_{i-1}, ..., X_{i-1} and are independent of each other. The units for rainfall throughout take 0.01 inch/hour to be 1.0. The term 0.5 in Eq. 5-2 is present to make the minimum generated rainfall 0.005 inch/hour. Intensities less than this are reported as zero in the hourly rainfall records.

Equation 5-1 represents the background process as an autoregressive process with dependence on n previous values. Such processes are often denoted as AR(n) in the literature with AR indicating an autoregressive process and the n denoting the number of previous values involved in the relationship. Salas et al [1980] and Bras and Rodriquez-Iturbe [1985] present good summary discussions of AR processes. The AR(n) process produces values which are normally distributed. We further require that the background process in this application also produce standard normal values so that the mean is zero and the variance is unity.

Equation 5-2 transforms those values of the background process which exceed the threshold such that they look like rainfall. This transformation is given by the sum of two random variables: one a power transform of the background process value and the other a power transform of an independent random variable distributed identically to the truncated background process. This approach to the transformation will be discussed after the parameter estimation for the background process is discussed.

Estimation of Parameters

The parameters for the background process to be

BEST COPY AVAILABLE

determined are: h, the truncation level, n, the number of lagged values in the autoregressive process, and $D_0,\ldots,D_n$, the coefficients in the autoregressive process. The truncation level, h, is defined by the requirement that the proportion of time it is raining be mimicked by the model. We then get

$$P(R_i > 0) = P(X_i > h) = 1 - G(h) \qquad (5\text{-}4)$$

where G(x) = cumulative distribution function (CDF) of the standard normal random variable. The probability of rainfall in any hour is estimated from the observed data by computing the ratio of the number of hours with rain to the number of hours in the season.

The probability of a storm and the distribution of the length of a storm is used to determine the remaining parameters for the background process. A storm is defined as a consecutive sequence of hours with rain. The background process indicates the beginning of a storm whenever it is true that $X_{i-1} < h$ and $X_i > h$. This pattern is called an upcrossing of the level h or just an upcrossing when the level is implied by the context.

The observed probability of a storm of any length in a season is given by the ratio of the number of storms observed divided by the number of hours in the season. The probability of a storm for the background process is

$$P(X_{i-1} < h \text{ and } X_i > h) = L(-h,h;-r) \qquad (5\text{-}5)$$

where

$$L(h,h;r) = \int_h^\infty dx \int_h^\infty f(x,y)\, dy \qquad (5\text{-}6)$$

f(x,y) = bivariate normal density =

$$= \frac{1}{2\pi\sqrt{1 - r^2}} \exp\left[- \frac{x^2 - 2rxy + y^2}{2(1 - r^2)} \right] \qquad (5\text{-}7)$$

and r = lag-1 autocorrelation in the background process. Note that fixing r fixes the probability of a storm and any additional lagged correlations have no effect.

Equation 5-5 was solved for r given the value of h from Equation 5-4 and the observed probability of a storm for each

BEST COPY AVAILABLE

season. Table 5-1 gives the values of h and r for each of the seasons at Redwood Valley together with the observed probability of rainfall and a storm.

TABLE 5-1

Truncation Level and Lag-1 Correlation at Redwood Valley

Season	Probability of Rainfall	Probability of Storm	Truncation Level	Lag-1 Correlation
1	0.0422	0.0109	1.726	0.953
2	0.0716	0.0149	1.464	0.963
3	0.1257	0.0235	1.147	0.959
4	0.1618	0.0287	0.987	0.957
5	0.1543	0.0308	1.018	0.947
6	0.1820	0.0371	0.907	0.938
7	0.1598	0.0293	0.995	0.954
8	0.1276	0.0297	1.138	0.936
9	0.1137	0.0276	1.207	0.935
10	0.0636	0.0163	1.525	0.946
11	0.0570	0.0161	1.581	0.937
12	0.0211	0.0068	2.031	0.942
13	0.0042	0.0014	2.637	0.954
14	0.0121	0.0028	2.255	0.975

An AR(1) process for the background process is defined once h and r are estimated. This simple process is given by

$$X_i = rX_{i-1} + \sqrt{1 - r^2}\, Z_i \qquad (5\text{-}8)$$

background process is fixed by the truncation level, h, and the lag-1 autocorrelation, r. The adequacy of the AR(1) process for the background process is determined by computing the storm length distribution for the observed rainfall and the storm length distribution implied by Eq. 5-8. If these distributions are close, then the AR(1) process is an adequate approximation to the background process.

The degree of closeness of the two distributions was measured using the maximum difference between their CDF's. This is the statistic used in the Kolmogorov-Smirnov test. The storm lengths were assumed to be independently distributed. The test is strictly applicable for continuous distributions but application to discrete distributions tends to be conservative. It is furthermore limited to completely specified distributions for which no parameters are estimated

BEST COPY AVAILABLE

from the data. Some parameters are estimated from the data in the current application and an adjustment discussed by Crutcher (1975) was used. The coefficient in the asymptotic formula for the 5% probability level is 1.36 when no parameters are estimated. Crutcher gives various coefficients depending on the distribution type and the nature of the parameter being estimated. None of these distributions are applicable so a value of 0.90 for the coefficient was used. This value is within 0.014 of the minimum value given by Crutcher for the 5% probability level.

A storm of exactly m hours duration occurs whenever the following sequence of values in the background process occurs:

$$X_0 < h,\ X_1 > h, \ldots, X_m > h,\ X_{m+1} < h \qquad (5\text{-}9)$$

Thus m + 2 values of the background process are involved in the definition of a storm of length m-hours. Let $x = \{x_0, x_1 \ldots, x_m, x_{m+1}\}$ = vector of values defining a storm of length m and $f(x)$ = joint density function of the m+2 variables. Then the probability that the sequence defining a storm of exactly length m-hours is

$$P(L=m) = \int_{-\infty}^{h}\int_{h}^{\infty}\cdots\int_{h}^{\infty}\int_{-\infty}^{h} f(x)\ dx \qquad (5\text{-}10)$$

A storm must occur before it makes sense to talk about the distribution of storm lengths. Therefore, the probability of a storm of length m given that a storm has occurred becomes

$$P(L=m|\ \text{storm}) = \frac{P(L=n)}{P(\text{storm})} \qquad (5\text{-}11)$$

The multiple integral in Eq. 5-10 cannot be evaluated in general by analytic or numerical integration. However, we are not concerned with the general case; only the specific case of an autoregressive background process with small lag. If the background process is AR(1) then the integrand can be written as

$$f(x_{m+1}|x_m)\ f(x_m|x_{m-1})\ \ldots\ f(x_1|x_0)\ f(x_0) \qquad (5\text{-}12)$$

and if the background process is AR(2)

$$f(x_{m+1}|x_m,x_{m-1})\ f(x_m|x_{m-1},x_{m-2})\ \ldots$$
$$f(x_2|x_1,x_0)\ f(x_1|x_0)\ f(x_0) \qquad (5\text{-}13)$$

BEST COPY AVAILABLE

and so forth for general lag n. The number of factors involved is always m+2 so that the maximum number of conditioning variables is m+1. For example, the most complex integrand for a 1-hour storm is

$$f(x_2|x_1,x_0)\ f(x_1|x_0)\ f(x_0) \qquad (5\text{-}14)$$

no matter what the lag, n, is. Thus, only the parameters which relate these three variables can affect the probability of a 1-hour storm. Higher order autocorrelations have no impact and the probability of a 1-hour storm is fixed once the lag-1 and lag-2 autocorrelations have been fixed. This argument carries forward to the probability of a storm of length m. This probability is fixed once the lag-1 through lag-(m+1) autocorrelations are fixed.

In principle, any storm length distribution could be fitted by appropriately defining the autocorrelation structure. This is in principle only because the autocorrelation structure for an AR process must satisfy certain restrictions. In this context the variance of the lag-0 variable conditional on the non-zero lagged variables must be positive. See Appendix A for further discussion.

The special structure of the integrand in Eq. 5-10 makes it possible to evaluate the multiple integrals using iterated numerical integration if the lag, n, is small. The equations for the general case are complex and the notation obscures the simple relationships involved. Therefore, a few specific examples will be given to illustrate the approach. The one hour storm probability is evaluated as follows:

$$g_2(x_1,x_0) = \int_{-\infty}^{h} f(x_2|x_1,x_0)\ dx_2 \qquad (5\text{-}15)$$

$$g_1(x_0) = \int_{h}^{\infty} g_2(x_1,x_0)\ f(x_1|x_0)\ dx_1 \qquad (5\text{-}16)$$

$$P(L=1) = \int_{-\infty}^{h} g_1(x_0) f(x_0) dx \qquad (5\text{-}17)$$

BEST COPY AVAILABLE

The two-hour storm probability is computed as

$$g_3(x_2,x_1) = \int_{-\infty}^{h} f(x_3|x_2,x_1)\, dx_3 \qquad (5\text{-}18)$$

$$g_2(x_1,x_0) = \int_{h}^{\infty} g_3(x_2,x_1) f(x_2|x_1,x_0)\, dx2 \qquad (5\text{-}19)$$

$$g_1(x_0) = \int_{h}^{\infty} g_2(x_1,x_0) f(x_1|x_0)\, dx1 \qquad (5\text{-}20)$$

$$P(L=2) = \int_{-\infty}^{h} g_1(x_0) f(x_0)\, dx_0 \qquad (5\text{-}21)$$

In each case f(|) is the conditional normal density function with one dependent variable and either one or two independent variables. The above equations simplify if the background process is AR(1). In this case the g functions are all functions of one variable and all the conditional densities have only one conditioning variable.

Each integral in the above equations must be evaluated numerically. Let us assume that a 15-point Gaussian integration rule is used with the limits of infinity being replaced by some finite limit which results in neglibible error(a value of 6 to 7 has worked well for the computation of storm length). Each of the g functions forms part of an integrand. Therefore. their values must be known at each point of the integration rule and thus g 3 and g 2 must be computed at 15 x 15 = 225 points. Each of these points represents a numerical integration itself using 15 points. Clearly the number of computations can become large rapidly. Great savings in computation can be realized if the values of f(|) are precomputed and saved. These values always occur at the same points and they are used tens of thousands of times if the CDF of the storm length distribution is to be computed to say 50-75 hours.

Comparison of the storm length distributions computed when the background process was AR(1) and the observed storm

BEST COPY AVAILABLE

length distribution revealed a consistent overestimation of the number of very short storms. Thus experiments with an AR(2) process were conducted. This process is

$$X_i = [r_1(1 - r_2)/(1 - r_1^2)]X_{i-1} + [(r_2 - r_1^2)/(1 - r_1^2)]X_{i-2} + \sqrt{[(1 - r_2)(2 - (1 - r_2)/(1 - r_1^2))]}\,z_i \qquad (5\text{-}22)$$

where r_1 = lag-1 autocorrelation and r_2 = lag-2 autocorrelation. If $r_2 = r_1^2$, then the lag-2 process becomes identical to the lag-1 process. Thus the first experiments with the AR(2) process were made with the value of r_2 only slightly different than r_1^2.

These experiments revealed that the storm length CDF is sensitive to small deviations in the value of r_2. A change of 20% in the probability of a 1-hour storm could result from a change of less than 2% in the value of r_2. Thus small changes in r_2 caused large changes in the probabilities of the very short storms. The effect attenuated rapidly so that the impact of the change was limited to storms less than 5 or 6 hours duration. Thus it was decided to use an AR(2) process as the background process.

The only free parameter was r_2 because h was set by the probability of rain in any given hour and r_1 was set by the probability of a storm. Thus r_2 was selected to minimize the maximum difference between the observed and simulated storm length distribution in each season. Table 5-2 gives a summary of the results for the station at Yorkville. Overall the resulting maximum difference was about one-half of the estimated Kolmogrorov-Smirnov test limit at the 5% level. Table 5-3 gives the first 25 hours of the storm length distribution for season 5 at Yorkville. In Table 5-3 the headings have the following meanings:

BEST COPY AVAILABLE

LENGTH - length of storm in hours.
SIMPMF - simulatated probability mass function.
OBSPMF - observed probability mass function for the storm length.
SIMCDF - simulated cumulative distribution function.
OBSCDF - observed cumulataive distribution function.
PMFDIF - difference between observed and simulated probability mass function.
CDFDIF - difference between observed and simulated cumulative distribution functions.

Table 5-2

Storm Length Distribution Results at Yorkville

Season	$r_2 - r_1^2$	Max. Diff.	K-S Bound
1	-0.0119	0.047	0.098
2	-0.0140	0.035	0.080
3	-0.0077	0.032	0.063
4	-0.0075	0.035	0.057
5	-0.0042	0.024	0.057
6	-0.0016	0.022	0.050
7	-0.0010	0.036	0.057
8	-0.0075	0.020	0.058
9	-0.0096	0.024	0.055
10	-0.0066	0.027	0.072
11	-0.0213	0.023	0.072
12	-0.0149	0.044	0.106
13	-0.0152	0.057	0.114
14	-0.0256	0.079	0.173

The background process is now fully specified with the truncation level, h, lag-1 correlation, r_1, and lag-2 correlation, r_2, known. The remaining parameters to determine are A_1, A_2, p, and q which relate to the distribution of the hourly rainfall amounts. The first transformation used did not involve the auxiliary process, Y. However, the resulting value of lag-1 correlation in the rainfall amounts was too large because the lag-1 correlation in the background process was nearly 1.0. The non-linear transformation did not reduce the background process correlation enough to approximate the observed lag-1 correlation in rainfall amounts.

Therefore a means of reducing the lag-1 correlation in the simulated hourly rainfall but yet maintaining its distribution had to be devised. The

BEST COPY AVAILABLE

Table 5-3

Storm Length Distribution in Season 5 at Yorkville

LENGTH	SIMPMF	OBSPMF	SIMCDF	OBSCDF	PMFDIF	CDFDIF
1	.3350	.3267	.3350	.3267	-.0083	-.0083
2	.1568	.1793	.4918	.5060	.0225	.0142
3	.0959	.0916	.5877	.5976	-.0043	.0099
4	.0663	.0398	.6541	.6375	-.0265	-.0166
5	.0493	.0438	.7033	.6813	-.0055	-.0221
6	.0384	.0438	.7417	.7251	.0054	-.0166
7	.0309	.0239	.7726	.7490	-.0070	-.0236
8	.0255	.0478	.7982	.7968	.0223	-.0014
9	.0215	.0159	.8196	.8127	-.0055	-.0069
10	.0183	.0199	.8380	.8327	.0016	-.0053
11	.0158	.0199	.8538	.8526	.0041	-.0012
12	.0138	.0199	.8676	.8725	.0061	.0049
13	.0122	.0040	.8798	.8765	-.0082	-.0033
14	.0108	.0199	.8905	.8964	.0092	.0059
15	.0096	.0080	.9001	.9044	-.0016	.0043
16	.0086	.0040	.9087	.9084	-.0046	-.0003
17	.0077	.0159	.9164	.9243	.0082	.0079
18	.0070	.0080	.9234	.9323	.0010	.0089
19	.0063	.0120	.9296	.9442	.0057	.0146
20	.0057	.0080	.9354	.9522	.0023	.0168
21	.0052	.0080	.9405	.9602	.0028	.0196
22	.0047	.0000	.9453	.9602	-.0047	.0149
23	.0043	.0040	.9496	.9641	-.0003	.0146
24	.0039	.0040	.9535	.9681	.0000	.0146
25	.0036	.0000	.9571	.9681	-.0036	.0110

approach, as given in Eq. 5-2, was to simulate the rainfall as the sum of two random variables. The first random variable was a power transform of the excess of the background process above the trunction level. The second random variable was the power transform of a truncated standard normal variable truncated at the truncation level used for the background process. This auxillary process would be independent of the background process.

The values of A_1, A_2, and p for a given value of q were computed such that the mean hourly rainfall, the variance of hourly rainfall, and the lag-1 auto-covariance of hourly rainfall were matched exactly. The value of the remaining parameter, q, was then varied to minimize the maximum difference between the simulated and observed CDF of hourly rainfall.

BEST COPY AVAILABLE

Rewrite Eq. 5-2 as

$$R_i = V_i + W_i + 0.5 \tag{5-23}$$

Then

$$E(R_i) = E(V_i) + E(W_i) + 0.5 \tag{5-24}$$

$$Var(R_i) = E(V_i^2) - E(V_i)^2 + E(W_i^2) - E(W_i)^2 \tag{5-25}$$

and

$$Cov(R_i, R_{i-1}) = E(V_i V_{i-1}) - [E(V_i | V_{i-1} > 0, V_i > 0)]^2 \tag{5-26}$$

where all expected values are conditional on their arguments being > 0. The expected value of V_i in Eq. 5-26 is conditional on both $V_i > 0$ and $V_{i-1} > 0$ to be consistent with the definition of covariance.

The expected values in Eqs. 5-24, 5-25, and 5-26 can be computed from the standard normal densities. Let

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \tag{5-27}$$

and recall that f(x,y) is defined above. Then define

$$Q(h) = \int_h^\infty f(x)\, dx \tag{5-28}$$

and recall the definition of L(h,h;r) from Eq.5-6 above. Now define an operator U as

$$U(A, P) = \frac{A}{Q(h)} \int_h^\infty (x - h)^P f(x)\, dx \tag{5-29}$$

Then the univariate expected values are

$$\begin{aligned} E(V_i) &= U(A_1, p) \\ E(W_i) &= U(A_2, q) \\ E(V_i^2) &= U(A_1^2, 2p) \\ E(Wi^2) &= U(A_1^2, 2q) \end{aligned} \tag{5-30}$$

The bivariate expected values in Eq. 5-26 are

$$E(V_i V_{i-1}) = \frac{A_1^2}{L(h,h;r)} \int_h^\infty \int_h^\infty (x-h)^p (y-h)^q f(x,y)\, dx dy \tag{5-31}$$

and

BEST COPY AVAILABLE

$$E(V_i|V_i>0,V_{i-1})0 = \frac{A_1}{L(h,h;r)} \int_h^\infty (x-h)^p f(x,y)\,dxdy \qquad (5\text{-}32)$$

For a given value of q, Eqs. 5-24 through 5-26 were solved using Newton's method.

Once a value for all four parameters was known, the CDF of simulated rainfall was computed. Each of the two random variables in Eq. 5-23 is a power transformed, truncated, standard normal random variable. The transformation is one-to-one so that the density fucntion of V or W can be easily found. This density function is

$$f_V(v) = \frac{(v/A)^{(1/p-1)}\ f[(v/A)^{(1/p)} + h]}{ApQ(h)} \qquad (5\text{-}33)$$

where f[]= the univariate standard normal density, h > 0, p > 1, A > 0, and 0 <= v <= infinity. This density function applies to both V and W. Integration of this density yields the CDF as

$$F_V(v) = \frac{G[(v/A)^{(1/p)} + h] - G(h)}{Q(h)} \qquad (5\text{-}34)$$

where G() is the CDF for the standard normal random variable. The random variables V and W are independent and the CDF of their sum is given by

$$F_{V+W}(z) = \int_{-\infty}^{\infty} F_V(z - x)\ dF_W(x) \qquad (5\text{-}35)$$

according to Parzen [1960. p. 404]. Substituting the appropriate values in Eq. 5-35 and changing the integration variable to eliminate the singularity in the integrand at the lower limit yields

BEST COPY AVAILABLE

$$F_{V+W}(z) = \int_{h}^{b(z)} \frac{G[g(z,u) + h] - G(h)}{Q(h)^2} f(u)\, du \qquad (5\text{-}36)$$

where

$$b(z) = [z/A]^{1/q)} + h) \qquad (5\text{-}36a)$$

and

$$g(z,u) = \frac{[z - A_2(u-h)^q]^{(1/p)}}{A_1} \qquad (5\text{-}36b)$$

This integral is easily evaluated numerically using Gaussian integration rules of modest order. The CDF of rainfall is then

$$F_R(z) = F_{V+W}(z - 0.5) \qquad (5\text{-}37)$$

where $z > 0.5$.

A simple iterative search for the minimum difference between the observed and simulated CDF of hourly rainfall was used to define q. Dennis and Schnabel [1981] was the source for search techniques and nonlinear equations used in this study. The mean, variance, and autocovariance of hourly rainfall was matched simultaneously as outlined above. Table 5-4 gives results for the parameters at Yorkville together with the minimum difference obtained for the CDF's.

The K-S bound in the table was computed from

$$0.90/\sqrt{n_0} \qquad (5\text{-}38)$$

where $n_0 = n_r/G(n_r, r)$, n_r = number of hourly rainfall observations used to compute the observed CDF; r = lag-1 auto-correlation computed from the n_r observations of hourly rainfall; and $G(n_r, r)$ = function which corrects for the bias in the sample variance computed from a sample from a AR(1) process. This last function is given by

$$G(n_r,r) = 1 + \frac{2r}{1 - r} - \frac{2r(1 - r^n)}{n(1 - r)^2} \qquad (5\text{-}39)$$

BEST COPY AVAILABLE

Table 5-4

Transformation Parameters for Hourly Rainfall at Yorkville

SEAS	A1	A2	P1	P2	Max. Diff.	K-S Bound
1	16.055590	4.215805	1.740959	3.310648	0.081	0.093
2	11.255770	4.949847	1.795683	2.635403	0.051	0.069
3	8.959695	3.812018	1.948270	2.612911	0.020	0.055
4	10.826290	.527061	1.786951	4.437170	0.030	0.063
5	9.899666	2.300382	1.964283	2.994715	0.024	0.059
6	9.309358	.773772	1.733824	3.910225	0.026	0.051
7	10.687230	1.463707	1.602713	[illegible].559542	0.041	0.051
8	9.933330	.668778	1.730105	4.341062	0.034	0.059
9	8.774565	.877008	1.622583	3.895993	0.069	0.056
10	9.771399	1.881202	1.502868	3.267103	0.052	0.070
11	8.809379	1.657147	1.632744	3.401754	0.056	0.074
12	4.952631	8.647191	2.711081	2.027593	0.038	0.095
13	9.134576	9.680257	2.078759	2.001139	0.032	0.090
14	26.979190	1.358420	1.611901	5.138366	0.058	0.215

This adjustment is in the correct direction. However, it must be viewed as approximate only. The use of an unadjusted sample size in the formula for the K-S bound would yield a bound which was unreasonably small because the sample is not independent. This bound was used for the lack of a better alternative as a guide to the magnitude of the difference between the observed and simulated CDF which would be obtained even if the sample came from the simulated CDF.

All but one season had a maximum difference less than the estimated K-S bound. At Redwood Valley all seasons were less than the K-S bound. One value out of 28 exceeding the K-S bound is very close to the expected value of 5%.

Multivariate Model

All the parameters at a single station have been estimated above. These are h, the truncation level; r_1, the lag-1 autocorrelation; r_2, the lag-2 autocorrelation and A_1, A_2, p, q, the transformation parameters for for hourly rainfall.

Aggregation of two or more stations to create a multivariate model requires estimation of the cross

BEST COPY AVAILABLE

correlations between each pair of stations for lag-0, lag-1 and lag-2. These correlations must be based on the joint occurrence of rainfall at both stations at the appropriate lag relationships. The matrix of correlation coefficients, assembled as outlined in Appendix A, must also yield a positive definite matrix. Initial efforts at computing these cross correlations were based on determining the cross correlation in the background process required to yield the observed cross correlation between rainfalls when each station's background process was transformed as discussed above. This approach did not yield a positive definite covariance matrix for the six variables involved (2 lag-0, 2 lag- 1, and 2 lag-2). Attempts were made to slightly adjust some of the cross- correlations but the interrelationships were too complex, and the approach was abandoned.

Thus an alternative approach was selected. In this approach the cross correlation in the background process is selected such that the probability of joint occurrence of rainfall at the station and lag combination is reproduced. This is consistent with the approach used for determining the lag-1 and lag-2 autocorrelations at each station in that the magnitude of the rainfall amount did not enter into the compuations. Estimates of joint occurrence probability were also much easier to compute and should have smaller sampling variability. This approach produced a positive definite covariance matrix in every season and therefore was the method used for estimating the cross correlations for the multivariate case.

Table 5-5 gives the three correlation matrices for Season 6. Other seasons showed a similar pattern. The subscript on the matrix gives the time relationship between the two stations. For example, the upper right element in M1 gives the lag-1 cross correlation in the background process required to duplicate the observed probability that rainfall at Redwood Valley and at Yorkville one hour previous to the hour at Redwood Valley will both be positive. Similarly the lower left element in M_2 gives the lag-2 cross correlation in the

BEST COPY AVAILABLE

background process required to duplicate the observed probability that the rainfall at Yorkville and at Redwood Valley two hours previous to the hour at Yorkville will both be positive. The diagonal elements of M_1 and M_2 give the lag-1 and lag-2 auto- correlations at the stations respectively.

Table 5-5

Correlation Matrices for Season 6

	M_0		M_1		M_2	
	RED	YOR	RED	YOR	PED	YOR
RED	1.000	0.891	0.938	0.884	0.865	0.860
YOR	0.891	1.000	0.882	0.946	0.857	0.894

These correlation matrices, plus the truncation levels, and the transformation parameters for each season define the characteristics of the simulated rainfall. The defining equation for the background process becomes:

$$X_i = D \begin{bmatrix} X_{i-1} \\ X_{i-2} \end{bmatrix} + L Z_i \qquad (5\text{-}40)$$

where X_i = 2-element column vector of background process values at each station at hour i; D = 2 x 4 matrix of parameters computed from M_0, M_1, M_2; L = 2 x 2 lower triangular matrix computed from M_0, M_1, and M_2; and Z_i is a 2-element vector containing independent standard normal random variables. The matrices D and L are computed as follows:

Form the following matrices:

$$S_{11} = M_0$$

$$S_{12} = \begin{bmatrix} M_1 & M_2 \\ M_0 & M_1 \end{bmatrix} \qquad (5\text{-}41)$$

$$S_{22} = \begin{bmatrix} M_0 & M_1 \\ M_1' & M_0 \end{bmatrix}$$

Then compute

$$D = S_{12} \; S_{22}^{-1} \qquad (5\text{-}42)$$

and

$$LL' = S_{11} - S_{12} \; S_{22}^{-1} \; S_{12}' \qquad (5\text{-}43)$$

BEST COPY AVAILABLE

Finally find L from LL' using the Crout factorization given in Appendix A. The above equations are based on conditional sampling from the multivariate normal distribution. See Appendix A for additional discussion of these equations as well as the details of sampling from the normal and truncated normal distributions.

Equation 5-40 was used to generate values of the background process for each hour of the simulation period. At each seasonal boundary the parameters were changed to those for the new season. Equations 5-2 and 5-3 were used to determine the rainfall at each station as for the single station model.

Initial trials of this method produced storm length distributions which differed from the distribution defined by the parameters h, r_1, and r_2 at each station. These differences were within sampling variations for Yorkville but were larger than would be expected from sampling variations at Redwood Valley. The source of the problem was traced to the distortion of the single station correlogram by the multivariate AR process. A correlogram gives the auto-correlation as a function of lag number for a random process. The correlogram is defined for all lags by the parameters of the process. Thus if the process is a univariate AR(2) with zero mean, the lag-1 and lag-2 autocorrelations define the parameters. Taking expectations of Eq. 5-1 after post multiplication by a scalar X_{i-j}, j = n+1.... yields a recurrence relationship for the autocorrelation of any lag j > n as follows

$$r_j = d_1 r_{j-1} + d_2 r_{j-2} + \ldots + d_n r_{j-n} \qquad (5\text{-}44)$$

Generalizing Eq. 5-44 to lag n and postmultiplying by vector X_{i-j} yields the matrix form of Eq. 5-43

$$M_j = D \begin{bmatrix} M_{j-1} \\ M_{j-2} \\ \cdot \\ \cdot \\ \cdot \\ M_{j-n} \end{bmatrix} \qquad (5\text{-}45)$$

BEST COPY AVAILABLE

where M_j= correlation matrix at lag j, j > n, D = m x nm matrix, and m = number of stations in the multivariate process. In the multivariate case the diagonal elements of M_j give the auto-correlation values at lag j for each of the stations.

Table 5-6 gives the following values of the correlogram:

AR2S	= AR(2)	single station case
AR2M	= AR(2)	multivariate case
AR3M	= AR(3)	multivariate case forcing agreement with AR2S up to lag 3.
AR4M	= AR(4)	multivariate case forcing agreement with AR2S up to lag 4.
AR3M - AR2S =		difference between AR3M and AR2S.
AR4M - AR2S =		difference between AR4M and AR2S.

Table 5-6

Summary of Correlograms for Season 6

REDWOOD VALLEY

LAG	AR2S	AR2M	AR3M	AR4M	AR3M-AR2S	AR4M-AR2S
1	.938	.938	.938	.938	.000	.000
2	.865	.865	.865	.865	.000	.000
3	.796	.806	.796	.796	.000	.000
4	.731	.761	.741	.731	.010	.000
5	.672	.722	.698	.681	.026	.009
6	.618	.687	.662	.641	.044	.023
7	.568	.654	.628	.606	.060	.038
8	.522	.622	.596	.573	.073	.051
9	.480	.592	.564	.542	.084	.062
10	.442	.563	.533	.510	.091	.069

YORKVILLE

LAG	AR2S	AR2M	AR3M	AR4M	AR3M-AR2S	AR4M-AR2S
1	.946	.946	.946	.946	.000	.000
2	.894	.894	.894	.894	.000	.000
3	.844	.849	.844	.844	.000	.000
4	.797	.807	.800	.797	.003	.000
5	.753	.768	.759	.754	.005	.001
6	.711	.731	.719	.712	.007	.001
7	.672	.695	.680	.672	.008	.000
8	.635	.661	.643	.632	.008	-.003
9	.600	.629	.607	.594	.008	-.006
10	.566	.598	.574	.557	.007	-.009

Forcing agreement at lags greater than 2 was accomplished by computing the lag-3 and lag-4 correlation matrices using Eq. 5-45 and then changing the diagonal elements to agree

BEST COPY AVAILABLE

exactly with the lag-3 and lag-4 correlation values for the AR(2) single station case. This procedure always resulted in positive definite matrices so that the simulation parameters could be computed. The original discrepancy in correlograms at Yorkville (difference between columns headed AR2S and AR2M) was much smaller than the discrepancy at Redwood Valley. The discussion above established that the storm length distribution is defined by the correlogram. This is the reason for the larger sampling variation in storm length distribution found at Redwood Valley.

Limited studies were made to try to determine why the discrepancy was larger at Redwood Valley than at Yorkville. No definitive conclusions were reached but it seems likely that the asymmetrical discrepancy in storm length was caused by the difference in the lag-1 and lag-2 autocorrelations at the stations. These autocorrelations were always larger at Yorkville than at Redwood Valley. The lagged cross correlations did not seem to be involved because they do not differ much in some of the seasons. The exact cause of the difference in the discrepancy at the two stations need not be known with precision if a method can be found to make the discrepancy at each station as small as desired. That has been done.

The approach of forcing agreement with the correlogram defining the storm length distribution is quite effective. The discrepancy at lag-10 for Redwood Valley has been reduced by half with the addition of two lags. The discrepancy for Yorkville has been reduced to a neglibible value at the same time. However, the process was stopped at lag-4 because the agreement being forced was between the storm length distribution produced by the multivarite model and the storm length distribution produced by the single station model. The goal of the fitting process is to obtain close agreement between the storm length distribution produced by the background process and the observed storm length distribution.

The above results show that this can be done by fitting the observed storm length distribution by using lags greater

BEST COPY AVAILABLE

than 2 in the background process. Then the initial agreement between storm length distribution of the univariate and multivariate process will be much greater. In principle, the process could be carried out to a larger lag to force nearly exact agreement. There are good reasons for not doing so (mainly the sampling variation inherent in the observed storm length distributions) but the methodology exists to accomplish such agreement.

This process was not implemented in the current study because the estimation of the storm length distribution for lags greater than 2 would have required extensive recoding of computer programs and the use of a much larger computer. Testing showed that the addition of two more lags to force agreement with the AR2S results produced rainfall which was acceptable for current purposes. Any Phase II work on this project would, however, use the method outlined in the previous paragraph to maximize the agreement between the observed and the simulated storm length distribution.

Thus the final version of the background process used in the current study was an AR(4) process with the additional two lags being used to force closer agreement between the storm length distribution as defined by the single station model and the storm length distribution defined by the multivariate model.

Testing the model

The model must be tested to verify formal correctness of the numbers being produced and to establish the fit of the model to those characteristics of rainfall not directly specified by the parameter estimation process. For example, at no point in the estimation process was the distribution of the times between successive storms used. Also annual totals of rainfall did not enter the estimation process in any form. As mentioned in the introduction to this chapter, there are nearly an infinite number of characteristics of the hourly rainfall which could be checked for agreement. We must content ourselves with the more obvious characteristics and

BEST COPY AVAILABLE

with those which seem to be important for the application of the generated rainfall. The stochastic rainfall data are just an intermediate stage in the generation of streamflow and the study of flood frequency distributions. Many characteristics of the rainfall data may not significantly affect the final outcome of the study.

The following characteristics of the rainfall process were used explicitly in the estimation of parameters:

1. The probability of a storm
2. The probability of rain in any given hour
3. The average hourly intensity during storms
4. Variance of hourly intensity during storms
5. Auto-covariance at each station
6. The probability of joint occurrence at lags 0 through 2 between the two stations
7. The distribution of rainfall amounts during storms
8. The distribution of storm lengths

The generated rainfall should be within sampling variations of the above values as defined by the parameters. The distribution of storm length and the distribution of rainfall amounts were fitted only approximately. The other characteristics were matched exactly (within convergence tolerances) during the parameter estimation. This means that the parameters and the probability distributions implied by the parameters become the population values for the generated samples. Close agreement between the distribution of rainfall amounts as defined by the parameters and as estimated from a sample of generated rainfall data establishes the veracity of the estimation and generation procedures. The distribution of rainfall amounts as defined by the parameters is only an approximation to the distribution as defined by the observed rainfall. Thus there are two sources for discrepancies in the test of the distribution of rainfall amounts and of storm lengths. The first is the difference between the distribution as defined by the parameters and the distribution estimated from the observed rainfall. The second is the difference between the distribution as defined by the parameters and the distribution estimated from a sample of generated rainfall.

BEST COPY AVAILABLE

The acceptability of the first difference is established during the parameter fitting process and the acceptability of the second difference is established in the testing process.

comparisons reported here are based on a 1000-year sample of generated rainfall data. These data were analyzed with the same programs used to analyze the observed rainfall data. A two year "warmup" period was generated and the data discarded to eliminate any transients from assumed initial conditions.

The sampling variability for the distributions of hourly rainfall amount and of storm length was estimated using the Kolmogorov-Smirnov test used in the parameter estimation procedure. The same ad hoc adjustment for a correlated sample was made in the case of the rainfall data. In the current case no adjustment for estimated parameters was made because sample statistics were being compared with population values. The sampling variability for the other statistics is complicated by dependence in the hourly rainfall. The lack of a random sample makes the estimation of the sampling variability by classical means very difficult.

Hence, an alternative approach was selected. The values for the rainfall characteristics estimated from the observed rainfall are the population parameters for the generated rainfall data. They therefore become fixed constants. The corresponding values estimated from the generated rainfall data are random variables. For large sample sizes the resulting random variables for the first six characteristics are approximately normally distributed. Thus we have the population value for the characteristic and the sample value from a large sample.

We wish to test for equality between the population value and the sample value for each season and station. A simple graphical display of the relationship can be obtained by plotting the sample values against the population values in a scatter diagram. Figure 5-2 is an example of such a plot. Perfect equality would result if all the points in the scatter plot fell on a straight line through the origin with slope =

BEST COPY AVAILABLE

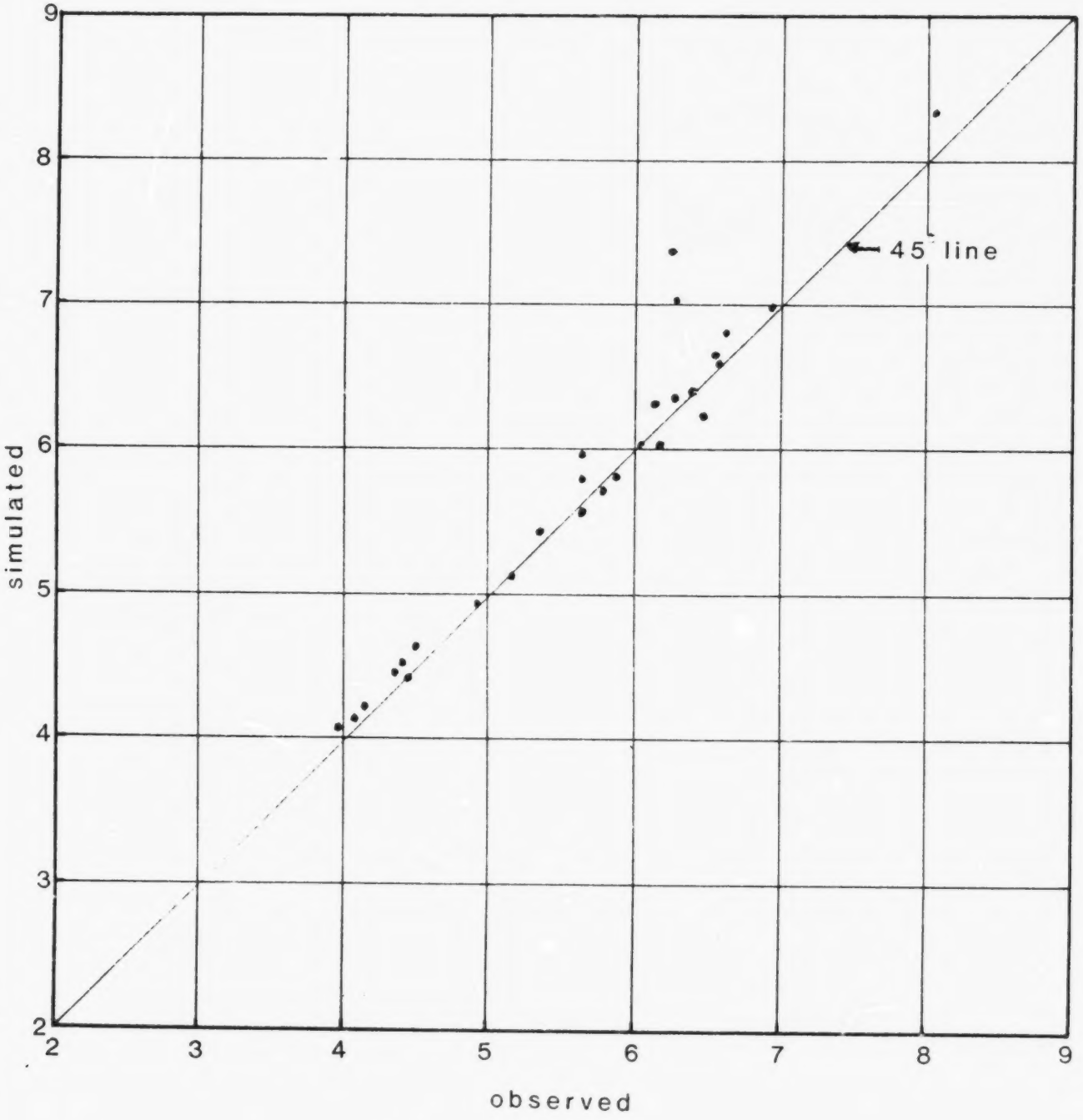

Fig. 5-2 Scatter plot of mean-hourly rainfall during storms

BEST COPY AVAILABLE

1. Perfect equality is rare and should not be expected. How close can we expect the scatter plot to be to the line of perfect equality? The approach taken here is to compute the least squares fit of a straight line with the population values taken as the independent variable. Then the sampling theory associated with least squares can be used.

The data for the scatter plots satisfies the assumptions for this theory except for the assumption of a constant variance about the regression line. It is likely that the sampling variance for each season and station pair for a given characteristic is slightly different. However, the nature of the variation is not known. If the nature of the variation were known, then the least squares estimation procedure could be modified to reflect the variation of the variance about the regression line. Johnston [1972] shows that using least squares when the variance is not constant results in inefficient estimates of the intercept and slope of the line. This means that the variance of these estimates would be smaller if the least squares estimation procedure were modified to explicitly account for the known variation of the variance.

Unmodified least squares results in estimates of the variance of the slope and intercept of the least squares straight line which are larger then necessary. Of course the reduction of the variance only results if the nature of the variation of the variance about the regression line is known. The examples given by Johnston [1972] show that the variance about the regression line must vary strongly with the independent variable before the variance of the least squares estimates are increased markedly. Quadratic variation increased the estimates of the variance of the slope by a factor of 1.8 and linear variation increased them by a factor of 1.2. Partial compensation for this effect can be effected by requiring that the desired slope of 1.0 and intercept of 0.0 be well away from the upper and lower confidence limits computed from unmodified least squares.

Table 5-7 gives the results for the first five

BEST COPY AVAILABLE

characteristics in the list above. The values for each characteristic for both stations were combined for the least squares computation. This gives a sample size of 28 for each least squares line for characteristics 1-5 and a sample size of 84 for the probability of joint occurrence because the values for all lags were combined into a single sample for the scatter plot.

TABLE 5-7

Test of Explicitly Mimicked Characteristics

Scatter Plot Slope and Intercept with 95% Confidence Intervals

	Slope			Intercept		
Characteristic	Value	Lower Limit	Upper Limit	Value	Lower Limit	Upper Limit
1. P(storm)	1.004	0.995	1.013	-2e-5	-2e-4	2e-4
2. P(rain)	1.0009	0.991	1.011	1e-4	-1e-3	1e-3
3. E(rain)	1.006	0.965	1.046	0.03	-0.20	0.26
4. Var(rain)	0.964	0.916	1.013	1.6	-0.70	3.9
5. Cov(rain)	1.007	0.950	1.064	-0.28	- 2.6	2.0
6. P(joint occurrence)	1.001	0.994	1.008	7e-5	-5e-4	7e-4

These results show that the characteristics are in fact closely mimicked by the parameter estimation and rainfall generation processes. The variance of hourly rainfall amounts is the only case in which the desired slope of 1.0 is perhaps too close to the confidence limit bound. These results are also an indirect verification of the random number generators used in the generation program.

Table 5-8 gives the results of the storm length distribution comparisons. This table gives the differences between the simulated CDF and the CDF defined by the parameters. As discussed above two lags were added to the background process to partially correct for the distortion of the single station correlogram in the multivariate model. The table gives the maximum difference at lag 10 or less between the single station correlogram with the AR(2) model used to estimate the storm length distribution and the correlogram obtained with the AR(4) multivariate background process used to generate the rainfall.

BEST COPY AVAILABLE

The storm length distribution results indicate that additional lags are needed to force the distribution at

Table 5-8

Storm Length Distribution Comparisons

	Redwood Valley			Yorkville		
Season	CDF Diff.	K-S Bound	Correlogram Difference	CDF Diff.	K-S Bound	Correlogram Difference
1	*0.061	0.018	0.208	0.015	0.019	0.015
2	*0.018	0.016	0.121	*0.021	0.016	-0.050
3	*0.045	0.013	0.200	0.007	0.012	-0.016
4	*0.017	0.012	0.066	0.004	0.011	0.008
5	0.008	0.011	0.058	0.005	0.011	-0.016
6	*0.017	0.010	0.069	0.007	0.010	-0.009
7	*0.017	0.011	0.064	0.005	0.011	-0.010
8	*0.028	0.011	0.098	0.008	0.011	-0.021
9	*0.024	0.012	0.069	*0.015	0.011	0.029
10	*0.019	0.015	0.186	0.012	0.014	-0.016
12	0.010	0.024	0.155	0.010	0.022	-0.028
13	0.013	0.023	0.040	0.006	0.022	-0.042
14	0.024	0.037	0.039	0.002	0.033	-0.029

Items marked with an * exceed the K-S bound.

Redwood Valley to conform to the distribution defined by the parameter estimation process. As discussed above, the addition of extra lags to force conformance of the correlograms was stopped at AR(4) because it became clear that the process worked well but that it would be better in subsequent applications of the method to fit the observed storm length distribution using the additional lags. This would lead to a closer fit to the observed distribution as well as to greater conformance of the correlograms from the multivariate model.

Table 5-9 gives the maximum CDF differences for the distribution of hourly rainfall. The Kolmogorov-Smirnov bound is given assuming the sample is random (independent bound) and assumming it is not random (dependent bound). The two bounds give an indication of the impact of the dependent sample adjustment described earlier.

Clearly the generation process mimics the distribution of rainfall as defined by the transformation parameters. In no case was the dependent bound exceeded. In many cases the CDF

difference was less than the independent bound even though the hourly rainfall is clearly positively correlated.

TABLE 5-9

Hourly Rainfall Amount Distribution Comparison

	Redwood Valley			Yorkville		
Season	CDF Diff.	Dep. Bound	Ind. Bound	CDF Diff.	Dep. Bound	Ind. Bound
1	0.014	0.021	0.009	0.010	0.018	0.008
2	0.008	0.016	0.007	0.008	0.015	0.007
3	0.005	0.012	0.006	0.005	0.011	0.005
4	0.007	0.010	0.005	0.008	0.010	0.004
5	0.002	0.011	0.005	0.003	0.010	0.005
6	0.002	0.010	0.005	0.002	0.009	0.004
7	0.009	0.010	0.005	0.005	0.009	0.004
8	0.005	0.012	0.005	0.008	0.011	0.005
9	0.006	0.013	0.006	0.007	0.011	0.005
10	0.007	0.017	0.008	0.007	0.014	0.006
11	0.006	0.017	0.008	0.006	0.015	0.007
12	0.003	0.029	0.013	0.004	0.026	0.012
13	0.005	0.029	0.013	0.003	0.024	0.011
14	0.007	0.038	0.017	0.011	0.033	0.015

The comparisons made so far have involved hourly rainfall characteristics which were used explicitly in the estimation of the parameters. The following comparisons involve characteristics which were not used in the estimation of the parameters.

Table 5-10 gives the scatter plot results for the mean and variance of the time between consecutive storms (called the interstorm length). Data are presented in this table with all seasons included and with seasons 13 and 14 deleted. Seasons 13 and 14 are the driest seasons and have very few data. Furthermore, the mean interstorm length in Season 13 is from 5 to 10 times longer than in the other seasons. The nature of least squares fitting is such that any large discrepancy in a value this far from the main body of points has strong influence on the results. The results for the mean value in Table 5-10 show this effect.

The all seasons results indicate that the interstorm length was not properly mimicked by the generation model. However, the deletion of Season 13 and 14 shows that the mean

BEST COPY AVAILABLE

was mimicked well. Examination of the values of each season showed that in most cases the difference in mean interstorm

TABLE 5-10

Interstorm Length Comparison

Scatter Plot Slope and Intercept with 95% Confidence Intervals

Characteristic	Slope			Intercept		
	Value	Lower Limit	Upper Limit	Value	Lower Limit	Upper Limit
All Seasons						
Mean	0.894	0.836	0.952	18.1	-0.12	23.8
Variance	0.483	0.445	0.521	7100	-3500	17,000
Seasons 13 & 14 Deleted						
Mean	0.988	0.902	1.074	1.8	-2.6	6.1
Variance	0.459	0.348	0.571	510	-1600	2600

length was less than 10%. However, the variance is clearly too small. Examination of the cumulative distribution functions of the interstorm length showed that the generated rainfall has too few interstorms in the length range of 1-3 hours.

One problem with the distribution of interstorm lengths is its highly skewed shape. There are many short interstorms but yet the range of interstorm length is very large. Table 5-11 gives an example of the observed and simulated interstorms.

The interstorm lengths have been edited to make the table of reasonable length. The maximum observed interstorm length in this season was 719 hours. The second longest was 322 hours. The data become very sparse at the longer interstorm lengths and many lengths are not observed in a record of 17 years. The 1000-year sample for the generated data has many more observations but there are still many values not observed. For example, an interstorm length of 300 hours was not observed in the 1000- year sample. The maximum length in the generated sample was 442 hours. The mean of the distribution is reproduced but the higher moments and therefore the distribution of interstorm lengths is not

BEST COPY AVAILABLE

reproduced.

TABLE 5-11

Interstorm Length Distribution for Season 6 at Redwood Valley

Length	Observed CDF	Simulated CDF
1	0.337	0.195
2	0.479	0.294
3	0.566	0.358
4	0.618	0.416
5	0.657	0.461
6	0.689	0.501
7	0.706	0.532
8	0.718	0.557
9	0.735	0.580
10	0.737	0.598
50	0.884	0.861
100	0.945	0.955
200		0.995
244	0.977	
303		0.9995
322	0.994	
442		1.0000
719	1.000	

The mean is reproduced because the parameter estimation process closely mimics the storm length distribution and the probability of a storm. Thus the mean interstorm length is also mimicked because the sum of the interstorm and storm lengths accounts for all possible hours. However, the variance and higher moments are not constrained and were not reproduced. The parameter fitting process focused exclusively on the storm length distribution and did not use or check the interstorm distribution at any point. Methods for improving the reproduction of the distribution of interstorm lengths will be discussed below in Discussion of Results.

Two additional checks were made of the generated rainfall data. Figures 6-5 and 6-6 show the comparison between the observed and generated annual totals. Figures 6-14 and 6-15 show the comparison of depth-duration frequency for the observed and generated hourly rainfall series at Yorkville and Redwood Valley.

Comparisons between the characteristics of simulated

BEST COPY AVAILABLE

streamflow using the observed hourly rainfall and the generated hourly rainfall are made in Chapter 6.

Discussion of Results

The testing results show that the stochastic rainfall model has many desirable features. The following characteristics of hourly rainfall can be used explicitly in the estimation of parameters and can be duplicated in the generated hourly rainfall data:

1. The probability of rain in any given hour.
2. The probability of a storm.
3. The mean hourly rainfall intensity.
4. The variance of the hourly rainfall intensity.
5. The autocovariance of the hourly rainfall intensity.
6. The probability of joint occurrence of rainfall between stations.
7. The distribution of storm length.

The following characteristics are not used in the estimation process but were none-the-less duplicated in the generated hourly rainfall data:

1. The mean value of the interstorm length.
2. The mean and variance of annual totals.
3. The CDF of rainfall amounts for various durations.

The variance and the distribution of interstorm length was not reproduced. Chapter 6 will show that this shortfall on the part of the current stochastic rainfall model did not appear to have an adverse effect on the results from the hydrologic simulation model. However, this may not be the case in all hydrologic regimes. It is, therefore, important to overcome this problem of the current model.

The problem is basically that the background process lends itself to fitting closely either the storm length distribution or the interstorm length distribution but not both. As pointed out in the parameter estimation section, fixing the lag-1 and lag-2 autocorrelation fixes the probability of a one-hour storm. It also fixes the probability of a one-hour interstorm. Experiments have shown

BEST COPY AVAILABLE

that agreement with both probabilities cannot be obtained in general because improving the approximation to one of these probabilities worsens the approximation to the other. Fundamentally an additional parameter or set of parameters must be introduced to enable matching both the storm and interstorm length distribution.

Clearly the current simplicity of the parameter estimation process should not be sacrificed. The ability to directly relate parameters of the background process to the probability of storms, the probability of rain, and the distribution of storm lengths are strong advantages of the current procedure. Any generalization should leave the current estimation and generation process intact.

The approach to be taken in future applications of the stochastic model will involve some additional concepts from Franz [1970]. He used the concept of storm lengths and interstorm lengths defined for the network as a whole and not for each station as in the current work. A storm for the network, called a network storm, was defined as a consecutive sequence of hours for which at least one station in the network had rain. A network interstorm was defined as a consecutive sequence of hours for which no station in the network had rainfall. In that study and in the following studies by Ott [1972] and Nasseri [1976] it was found that the sequence of network storms and interstorms was independently distributed. That is, knowledge of the current network storm length did not change the distribution for the length of the following network interstorm length.

The distortion of the interstorm length distribution can be eliminated by using two models in parallel. One of the models will be the same as the current model (network storm model) and the other will be a network interstorm model. The estimation of parameters for the network storm model will be refined to include more lagged correlations to improve the reproduction of the observed storm length distribution. This can be done to any desired level of accuracy given sufficient computer resources. The model performance will then reproduce

BEST COPY AVAILABLE

the major storm characteristics satisfactorily.

The distribution of interstorms will, however, not be duplicated and a separate model similar to the current model will be fitted to the network storm and interstorm sequence. The network storm and interstorm sequence will have a value of zero if no station in the network has rain and one if any station has rain in a given hour. This sequence then defines the network storms and interstorms. From this sequence the probability of a network storm, the probability of a hour with rain, etc. can be computed just as for the current model. The difference is that only the interstorms will be of interest and the autocorrelation values will be estimated to reproduce the observed distribution of network interstorms with no concern for the distribution of the lengths of the network storms.

It is clearly possible to fit this distribution closely by adding additional lagged autocorrelations to the background process. Furthermore, since the network storm and interstorm sequence is univarite, the network interstorm model will also be univariate and therefore simpler than the current stochastic model.

The process of modeling the rainfall will be as follows. Assume for descriptive purposes that the generated sequence will start with a network interstorm period. Thus, the network interstorm model will be invoked. However, it may not produce an hour with no rainfall. If it produces an hour with rain that hour is discarded, simulated time is not advanced and the value for the next hour is computed. The process continues until the interstorm model produces an interstorm. This network interstorm is then used in the generated series. Every network interstorm must be followed by a network storm. Thus the network storm model is invoked.

It also may not produce an hour which belongs to a network storm. In this case the hour is discarded, time is not advanced and the next hour is computed. The process is continued until a network storm is produced. This network storm is then used in the generated series. The models are

BEST COPY AVAILABLE

invoked in turn to produce the entity for which they have been fitted.

It might seem wasteful to discard all the invalid values produced by each of the models. However, there seems to be no way to avoid this because the background process for each model must be unbroken. The whole parameter estimation procedure is based on an unbroken sequence of background process values. Any break in either the univariate background process for the network interstorm model or the multivariate background process for the network storm model would invalidate the whole estimation procedure.

This approach to generation of rainfall assumes that the network storm and interstorm lengths are independent. All tests made to date indicate that this is the case. The presence of dependence in the network storm and interstorm sequence will only invalidate the approach if the dependence is strong. Weak dependence would likely not have much impact. Thus the approach has broad application.

This approach also assumes that the network storm model will be able to duplicate the characteristics of the interstorms contained within a network storm. These will be called network storm interstorms. Network storm interstorms are in part created by the definition of network storms and are in part short interstorms which occurred naturally at the station. Clearly the maximum length of the network storm interstorms is limited and most of them will be very short. The distribution of network storm interstorms was computed for both the 17-year period of observed record and for the 1000-year generated record. During the wet seasons (seasons 4 through 9) the maximum length of a network storm interstorm was generally less than 8-hours. The drier seasons had longer maxima. In all seasons but the dry season, more than one-half the network storm interstorms were one hour long. The maximum difference between the CDF of network storm interstorms from the observed and generated sample of data was always less than the Smirnov two-sample test statistic at the 5% level for all seasons. Thus the proposed network storm model will reproduce

BEST COPY AVAILABLE

the distribution of the network storm interstorms.

The combination of a network storm model and a network interstorm model as outlined above will permit the reproduction of the distribution of network interstorms as well as all the charactersitics currently being matched. The added computation is small because the network interstorm model is always univariate and only the background process for the sequence of network storm and interstorm lengths need be simulated. Rainfall amounts are not involved and therefore no transformation of the background process is required. As a result the additional effort will be less than that incurred in simulating an additional station.

BEST COPY AVAILABLE

VI. THE HYDROLOGIC SIMULATION

The simulation parameters (Table 4-2) resulting from the calibration were used to simulate 1000 years of synthetic streamflow at Cloverdale. For the purpose of this study it was assumed that Lake Mendocino does not exist and that there is no diversion from the Eel River to the Russian River. This assumption eliminates the need to make assumptions about the variation of flow diversions from day to day and also to assume some sort of operating policy for Lake Mendocino with respect to flood control and flow maintenance during the summer season. It is entirely feasible to simulate the reservoir operation and in some situations this would be necessary. In this study, simulation of the reservoir operation would only increase computer run times without any gains. In an actual application all components of the system would be simulated so that the resulting synthetic flows would reliably reflect the system as it is. Alternatively, it would be possible to simulate any assumed future system, if changes are expected.

Since the observed streamflow record covers a period from before the construction of Lake Mendocino to the present it is not in itself homogenous. To provide an "observed record" for comparison with the 1000 years of synthetic record, a 20-yr period of flow was simulated with the same assumptions and the historic rainfall. This "observed record" is free from errors resulting from the construction and operation of Lake Mendocino and any other land use changes that occurred during the period of record. Since it uses the same parameters as those used for the simulation of the 1000-yr sample, any differences between the short and the long sample reflect only the differences caused by the sample size and the differences between the historic and synthetic rainfall.

Figure 6-1 shows the flood frequency curve for Cloverdale based on the 1000-yr sample and the frequency curve calculated from the "observed simulation" using the extreme value type I distribution. A Log Pearson Type III curve is superimposed

BEST COPY AVAILABLE

BEST COPY AVAILABLE

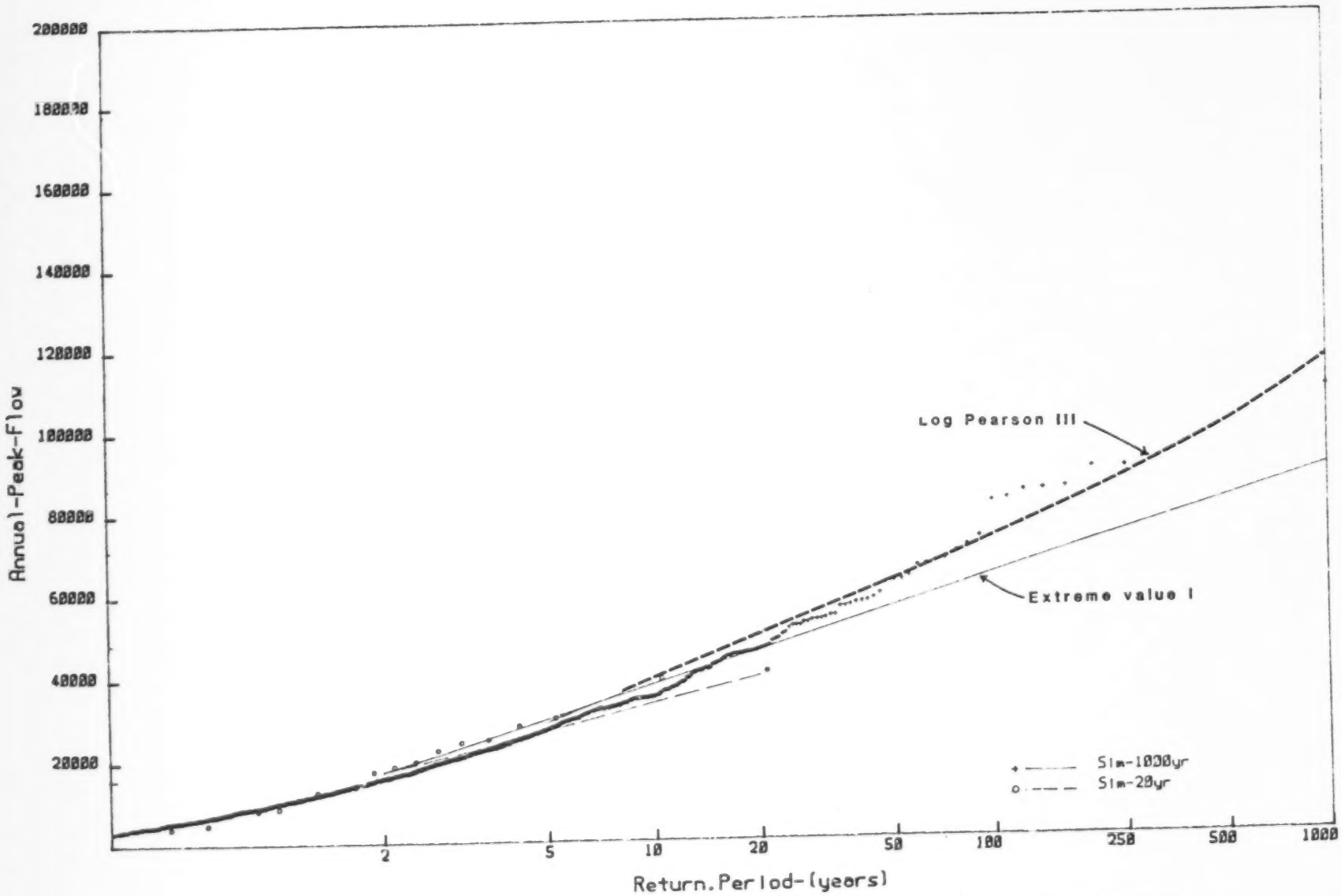

Fig. 6-1 Frequency curves for the 1000-yr and "observed" simulated samples using the Extreme value type I distribution. The log Pearson curve is plotted for comparison

for comparison. Both curves fit the sample data closely indicating that the variability induced by small samples has been largely eliminated. The log Pearson appears to be the best fit to the synthetic data. Applying the t-test, indicates that the mean and standard deviation of the data in the 1000-yr sample are the same as those in the historic sample at the 95 percent level.

Figure 6-2 shows frequency curves for the Russian River at Cloverdale based on two 500-yr periods taken from the synthetic record, ten 100-year, twenty 50-yr and fifty 20-year periods. The 500-yr samples depart only slightly from the 1000-yr curve. The smaller samples show increasing spread as the sample size decreases as be expected.

Figures 6-3 compares flow duration curves for Clovlerdale from the historic record and the 1000-yr sample. Again it is evident that there is excellent agreement over the entire range of flow.

Figures 6-4 through 6-6 compare the distribution of annual streamflow and rainfall for the upper Russian River catchment. Both factors are very well reproduced with the 1000-yr sample showing extremes, both high and low, which are slightly more extreme than the historic sample. This is what would be expected.

Figures 6-7 through 6-9 illustrate the effect of sample size on the mean, standard deviation, and coefficient of skew of flood peaks. For these figures the synthetic 1000-yr record was divided into samples if the indicated length and parameter values were computed for each sample. The largest and smallest values of each parameter were used to plot the curves. The skew coefficients were arithmetic.

The mean and standard deviation of the annual flow volume are shown on Figs. 6-10 and 6-11. On the basis of these comparisons it appears that a record of 1000 years define what may be called the true parameters for the streamflow. In some cases the 500-year samples agree with the 1000-yr sample. The general variation with sample size is what would be expected

BEST COPY AVAILABLE

BLANK PAGE

BLANK PAGE

BLANK PAGE

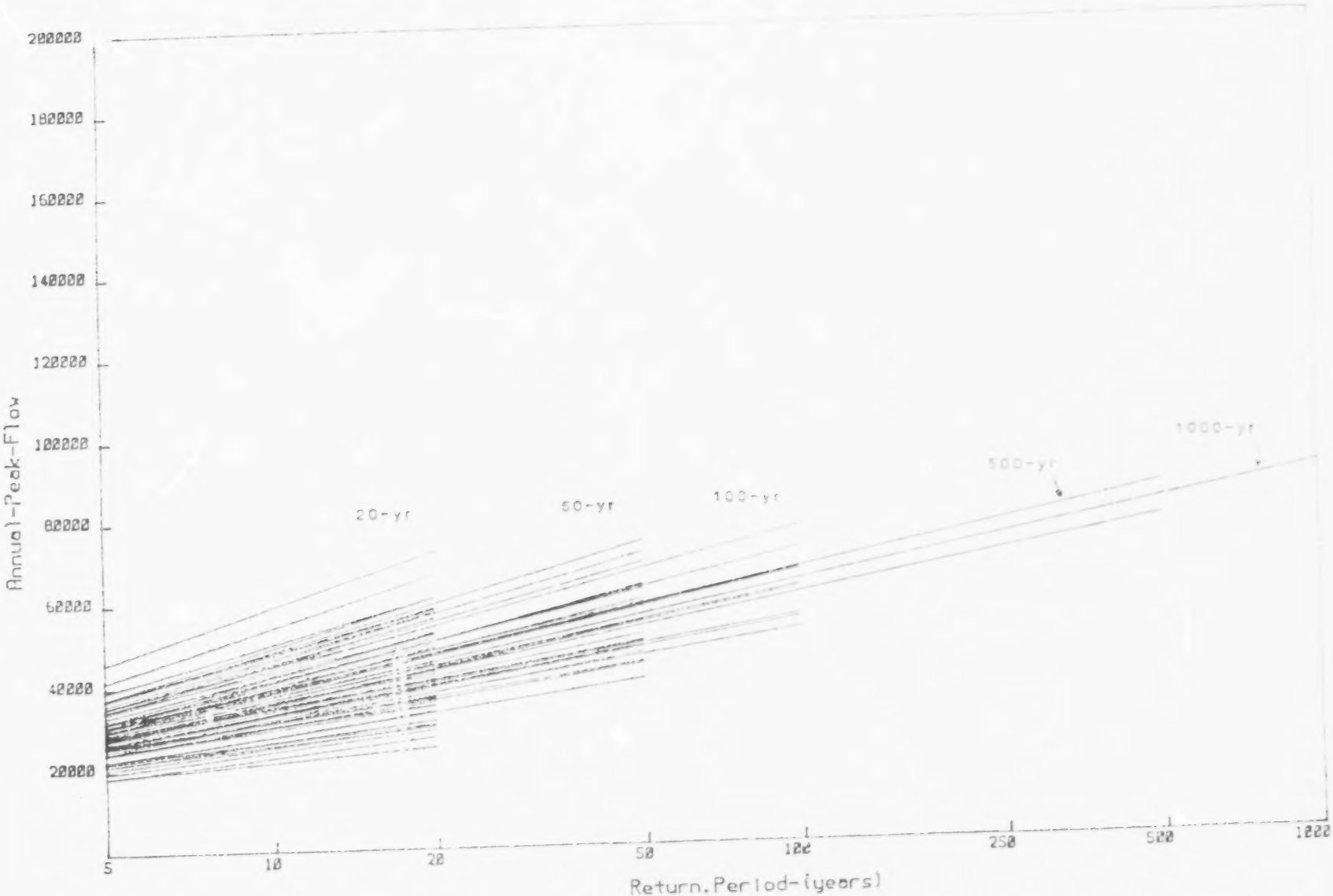

Fig. 6-2 Frequency curves from samples of 20, 50, 100, 500 and 1000 years drawn from the 1000-yr synthetic record

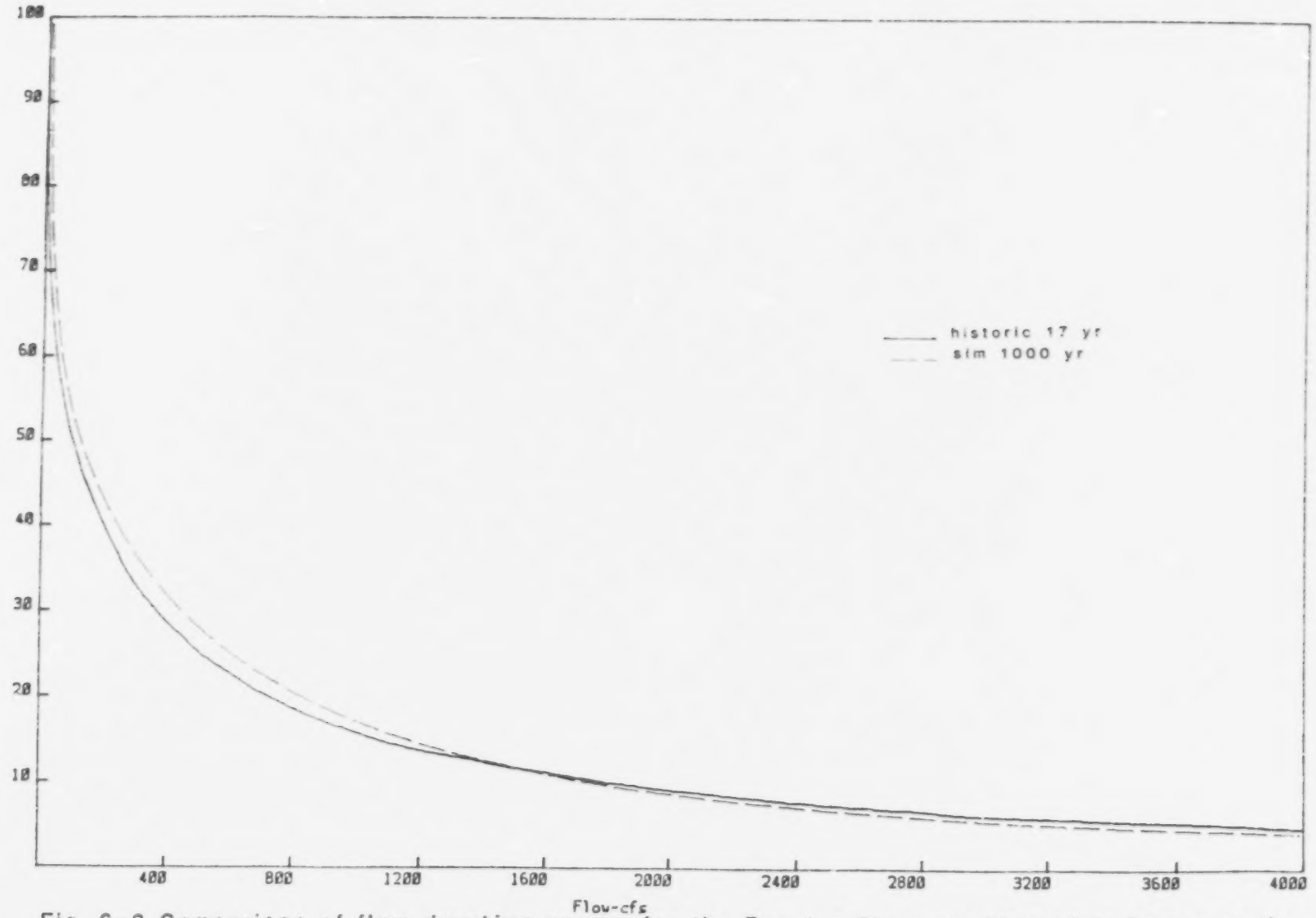

Fig. 6-3 Comparison of flow duration curves for the Russian River nr. Cloverdale based on the 1000-yr sample and the simulated historic record

BEST COPY AVAILABLE

89

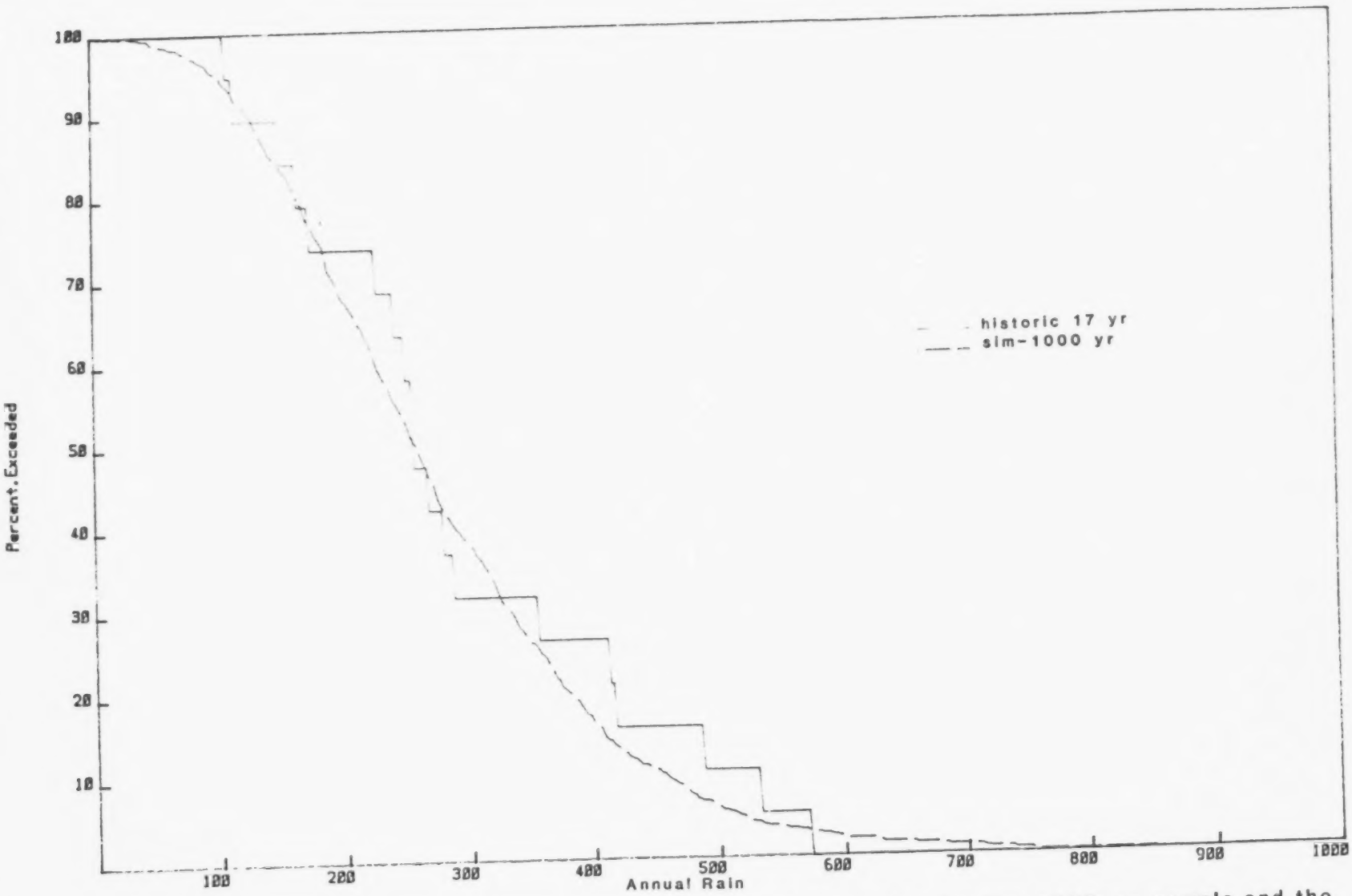

Fig. 6-4 Comparison of frequency of annual flows at Cloverdale for the 1000-yr sample and the 20-yr simulated historic record

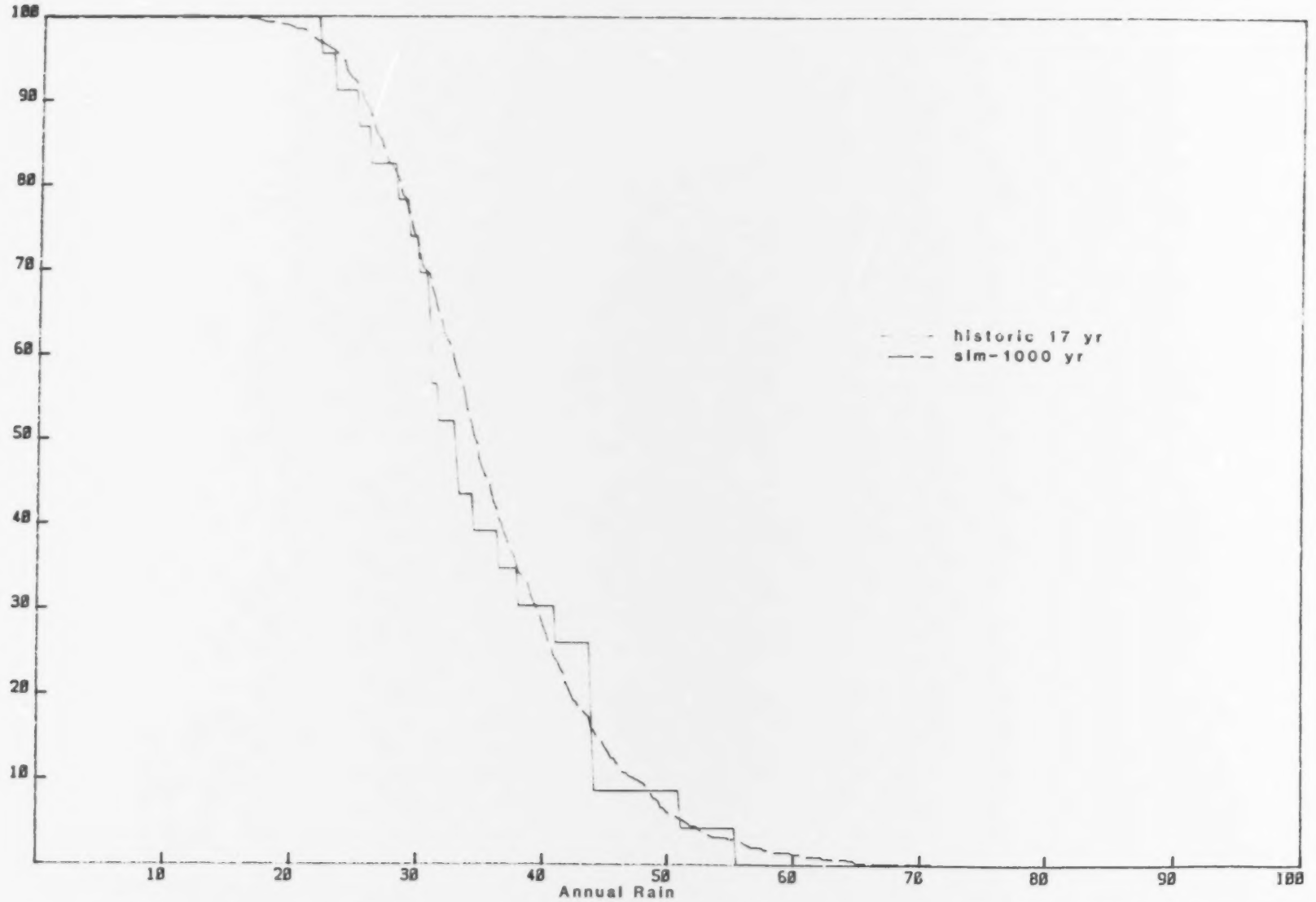

Fig. 6-5 Comparison of frequency of annual rainfalls at Redwood Valley for the 1000-yr stochastic generation and the historic record

BEST COPY AVAILABLE

BEST COPY AVAILABLE

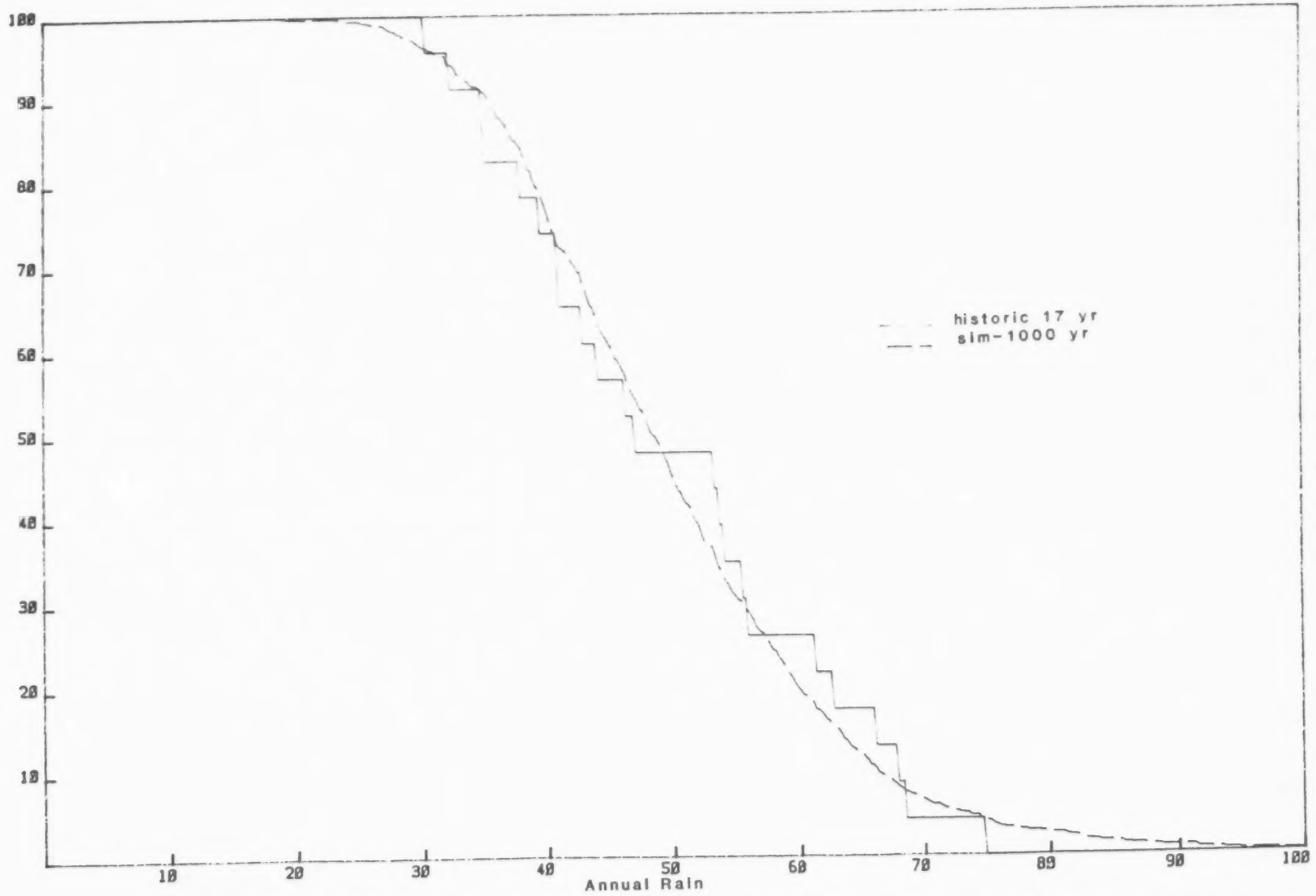

Fig. 6-6 Comparison of frequency of annual rainfalls at Yorkville from the 1000-yr stochastic generation and the historic record

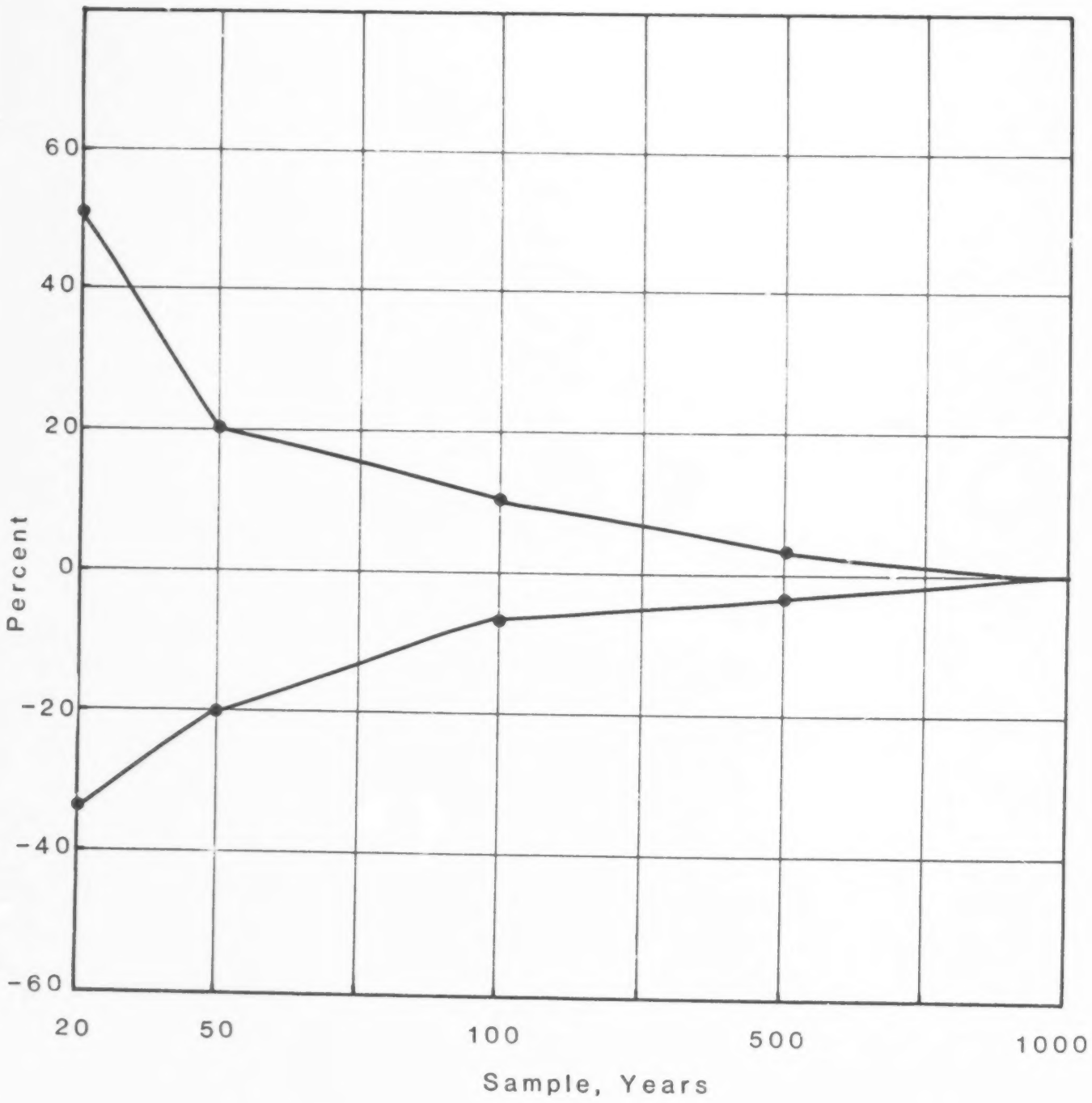

Fig. 6-7 Range of mean flood estimates at Cloverdale as a function of sample size

BEST COPY AVAILABLE

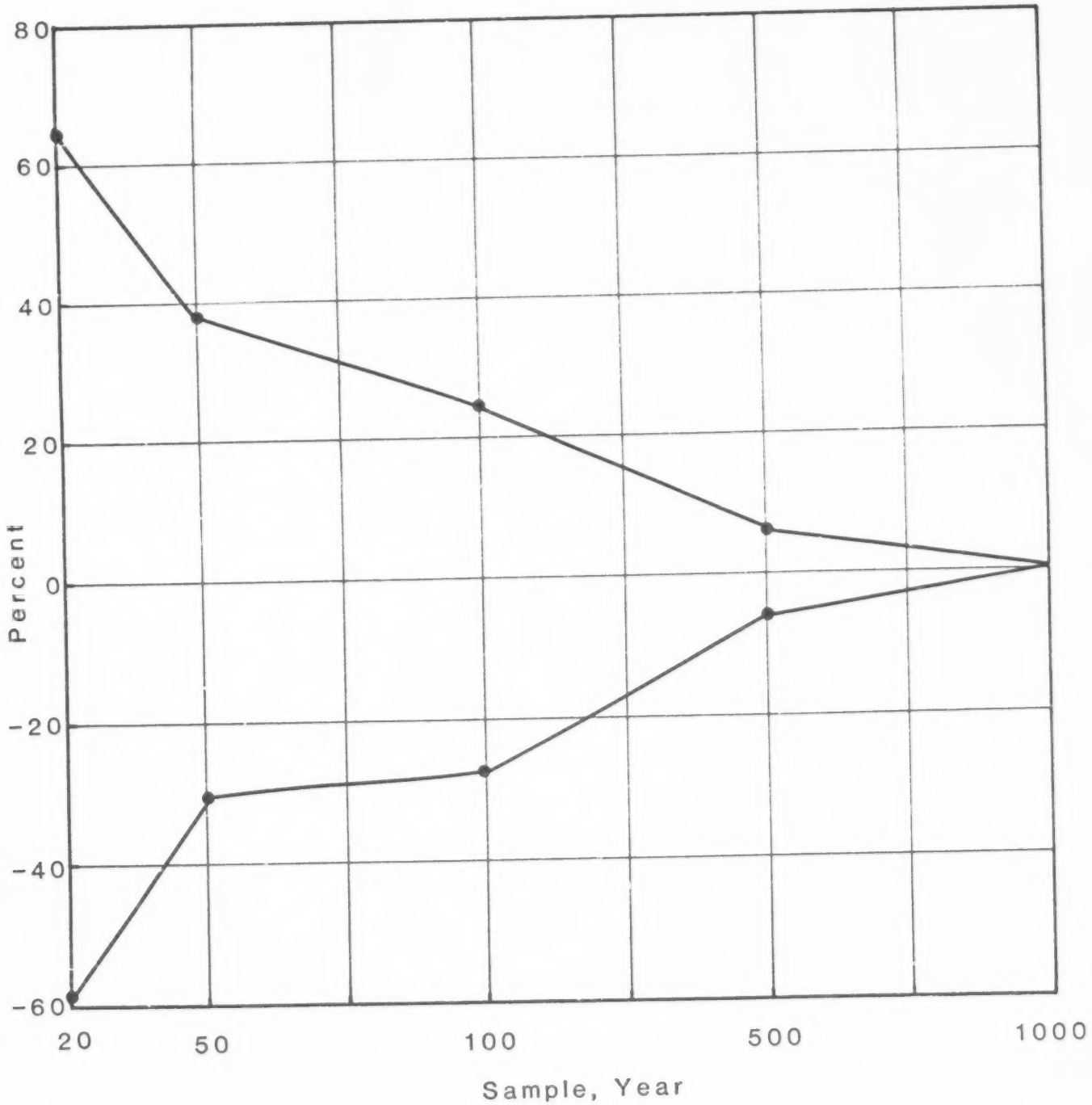

Fig. 6-8 Range of standard deviation of annual floods at Cloverdale as a function of sample size

BEST COPY AVAILABLE

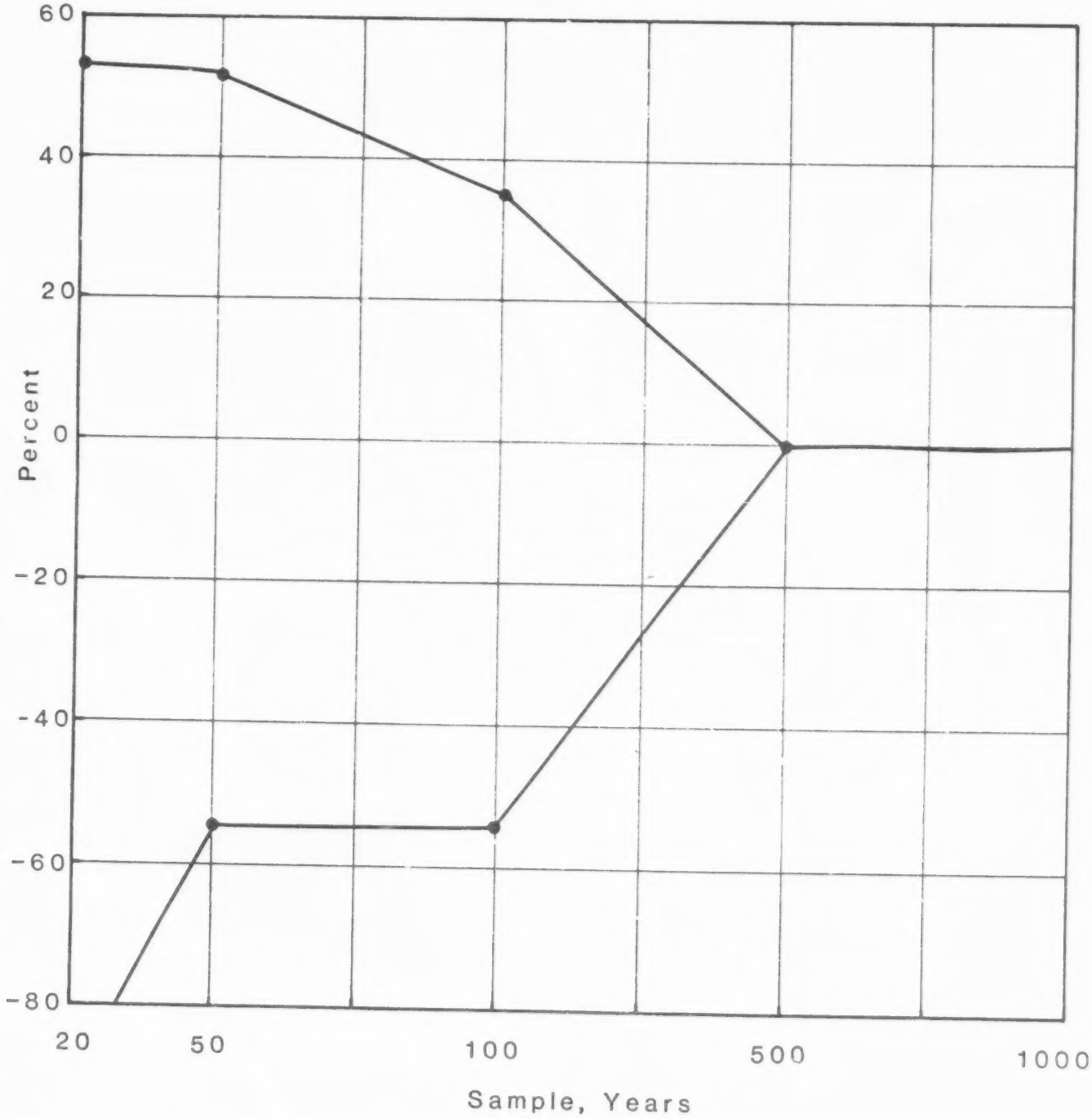

Fig. 6-9 Range of skew of annual flood peaks at Cloverdale as a function of sample size

BEST COPY AVAILABLE

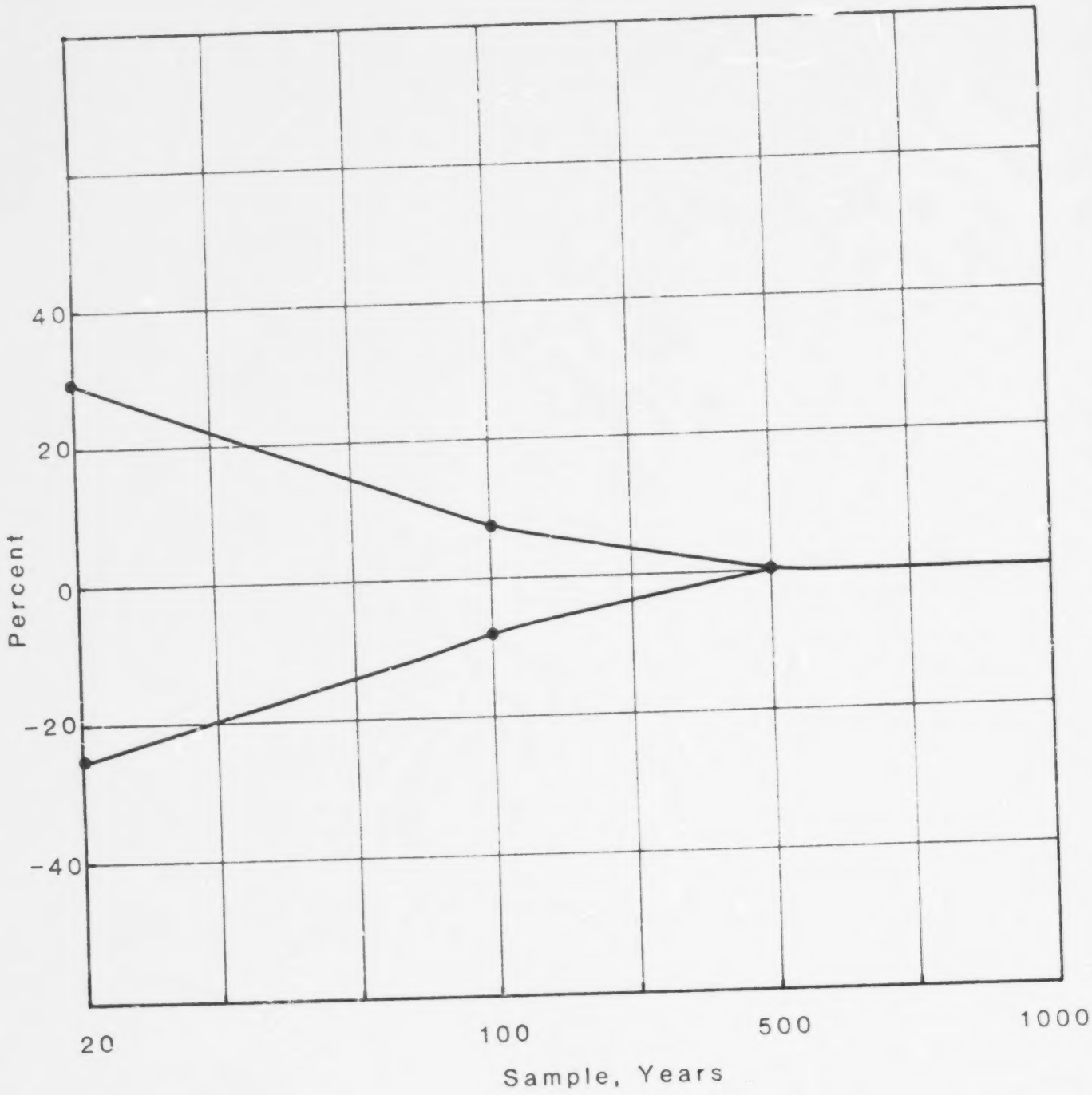

Fig. 6-10 Range of mean annual runoff at Cloverdale as a function of sample size

BEST COPY AVAILABLE

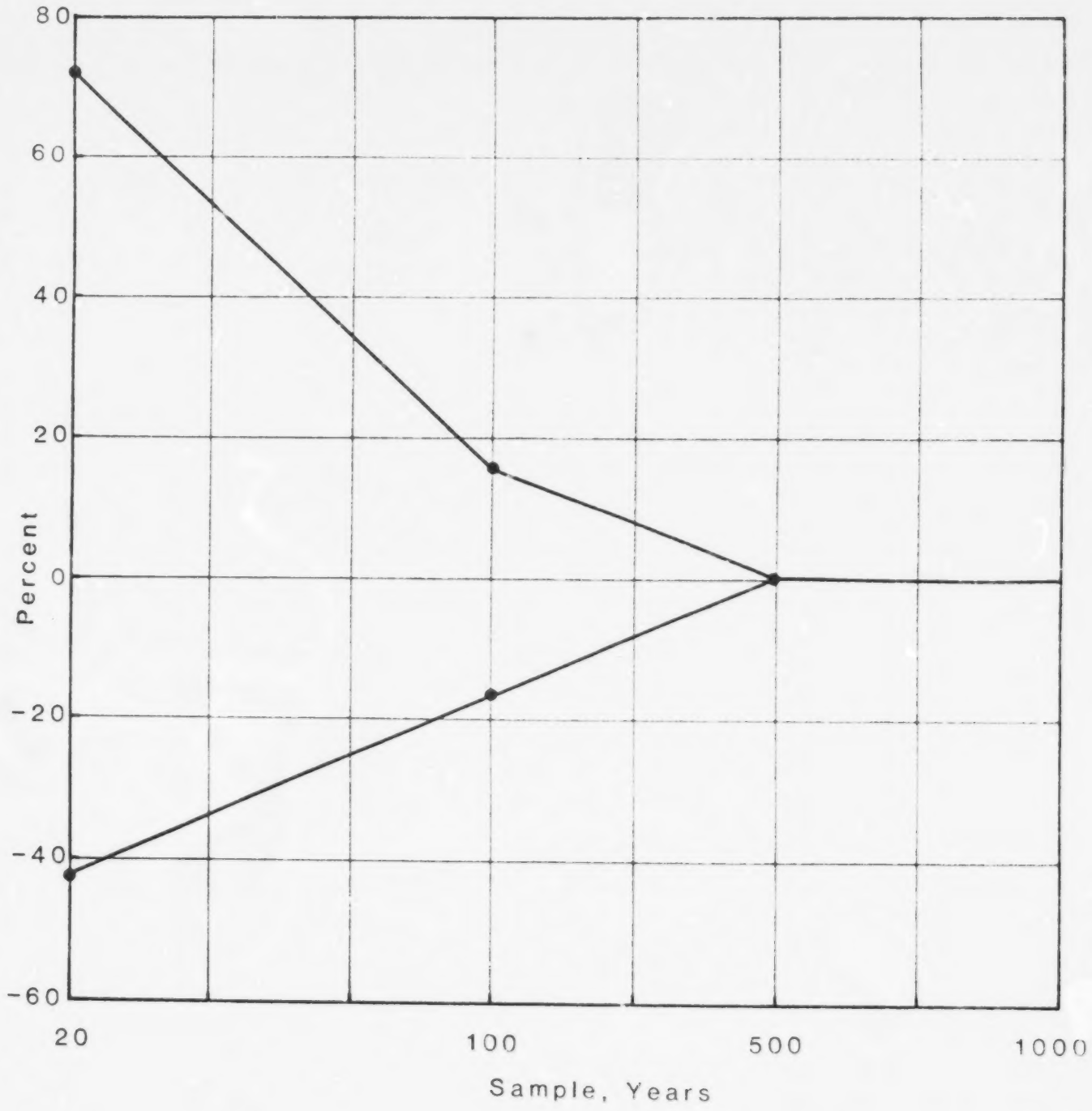

Fig. 6-11 Range of standard deviation of annual runoff at Cloverdale as a function of sample size

BEST COPY AVAILABLE

and agrees closely with the findings of Ott and Nasseri.

Figures 6-12 and 6-13 compare the variation of means and standard deviation with sample size for the rainfall data generated for Yorkville and Redwood Valley. These indicate that a 500-yr sample adequately defines these rainfall parameters. Data from Ott and Nasseri show very similar variation for both streamflow and rainfall.

Figures 6-14 and 6-15 compare values of rainfall intensity and probability for various durations at the two rainfall stations. These plots show excellent agreement between observed and stochastically generated rainfall. Again the concurrence of the curves can be described as excellent. They are in fact in surprisingly good agreement. These plots demonstrate that the rainfall reproduces the details of short duration rainfall very well.

Following the procedure specified in Hydrometeorological Report 36, [U.S. Weather Bureau, 1961] the probable maximum precipitation was determined for each of the segments of the Russian R. above Cloverdale. This rainfall was then converted to a flow hydrograph by simulation using the parameters of Table 4-2. The computed peak of 279,300 cfs (Fig. 6-16) is more than 5 times the highest observed peak at Cloverdale (55,000, Dec. 1964). The historic peak was reduced by the operation of Lake Mendocino so that this ratio should be somewhat lower. In comparison with the historic data, the probability of the computed PMF is quite small (about 1.2×10^{-10}) using the extreme value theory. In comparison with the 1000-yr synthetic sample, the probability of the PMF is still quite small.

The probability of the PMF is much lower than we had expected, but there is very little on which to base a judgement. It seems useless to speculate at this point. It may be noted that a 72-hr rainstorm with rainfall increasing monotonically up to a maximum and declining monotonically after the peak has never been observed in the records of the rain gages used in the study.

BEST COPY AVAILABLE

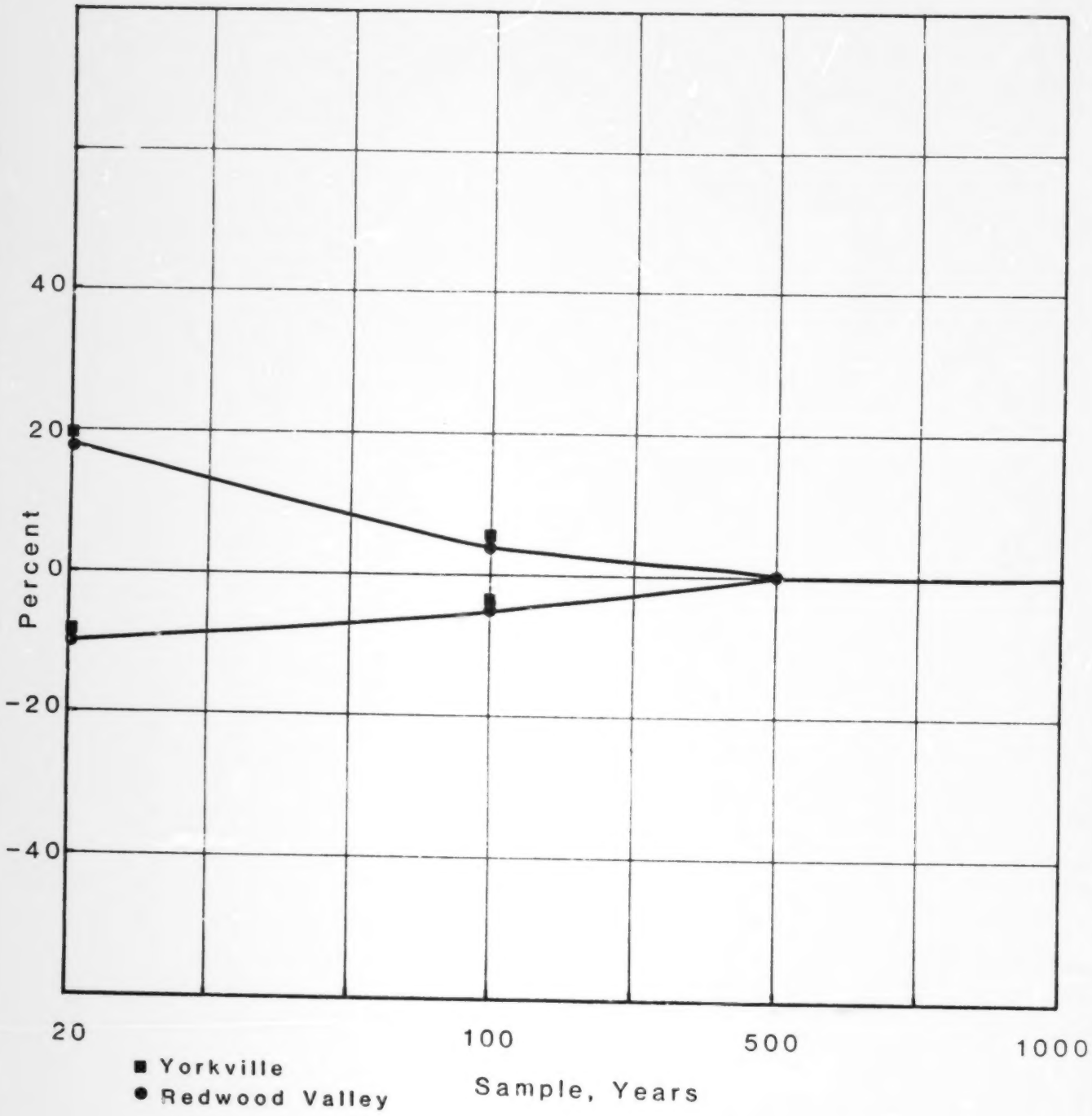

Fig. 6-12 Range of mean annual rainfall at Yorkville and Redwood Valley as a function of sample size

BEST COPY AVAILABLE

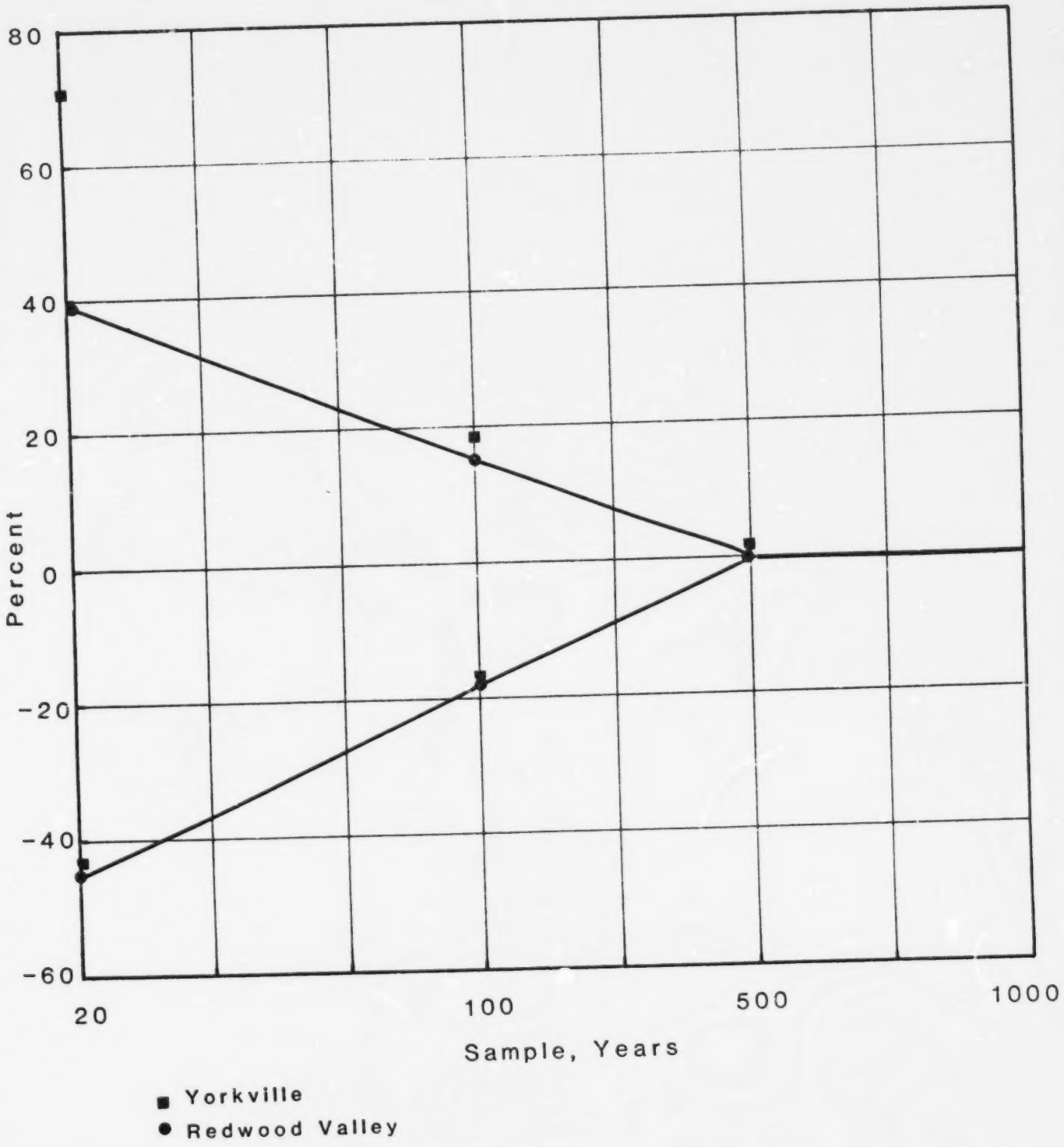

Fig. 6-13 Range of standard deviation of rainfall at Yorkville and Redwood Valley as a function of sample size

BEST COPY AVAILABLE

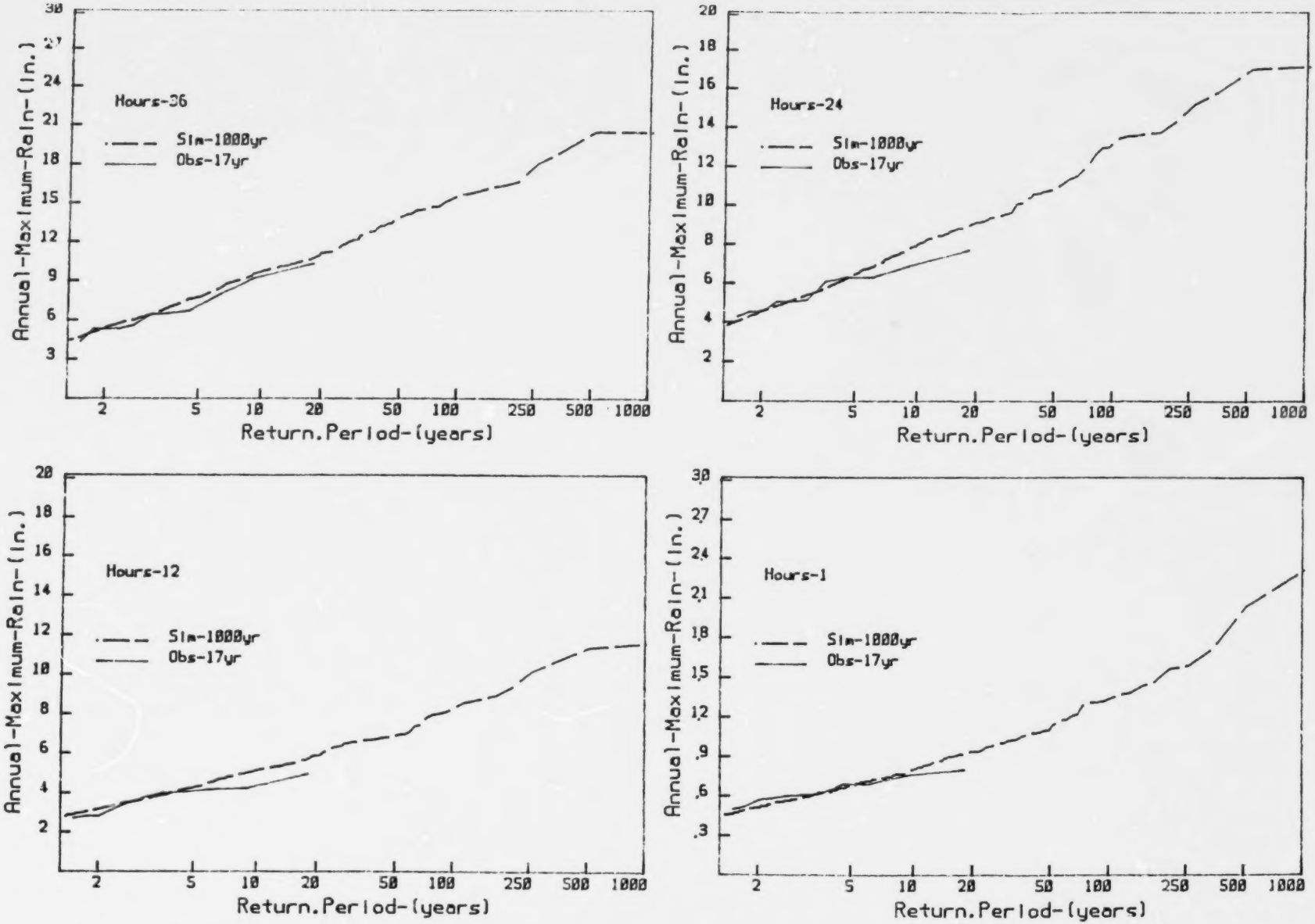

Fig. 6-14 Comparison of rainfall intensity and probability for various durations at Yorkville from observed and synthetic 1000-yr samples

BEST COPY AVAILABLE

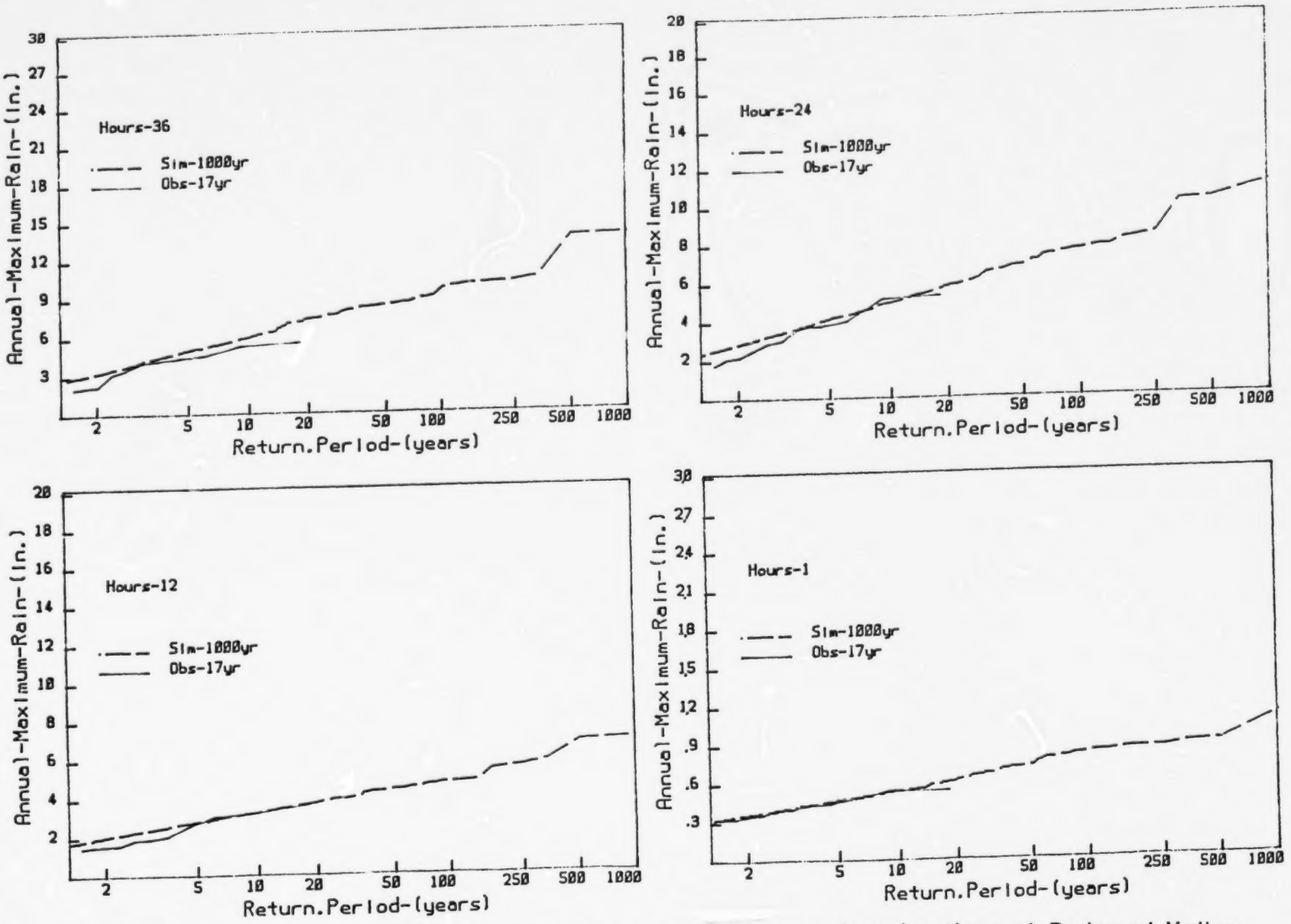

Fig. 6-15 Comparison of rainfall intensity and probability for various durations at Redwood Valley from observed and synthetic 1000-yr samples

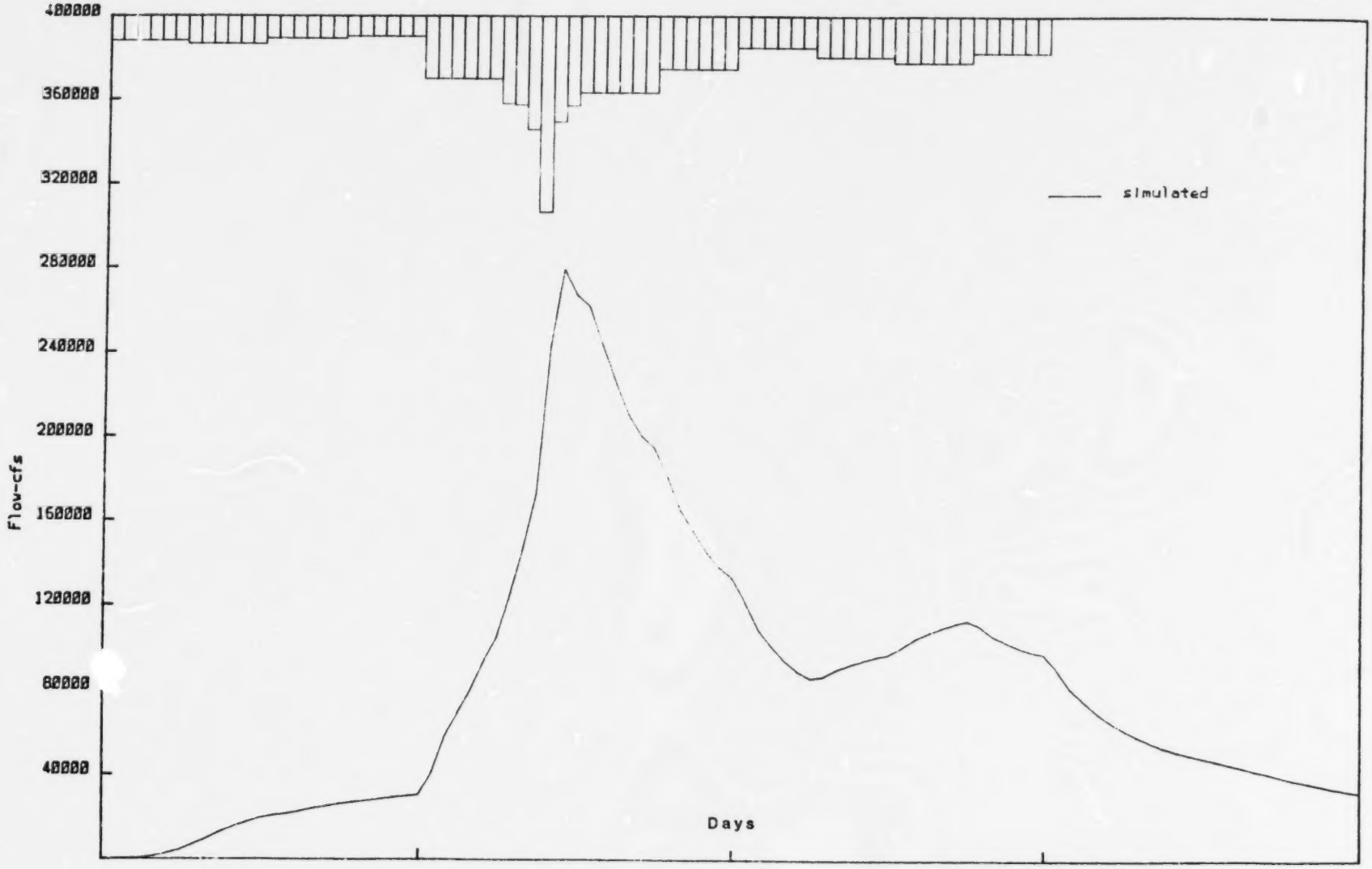

Fig. 6-16 Probable Maximum Flood, Russian R. nr Cloverdale

BEST COPY AVAILABLE

VII. CONCLUSIONS

o The goal of constructing a stochastic model of hourly rainfall data which will reproduce the major characteristics of the data has been reached.

o Explicit relationships exist between six major characteristics of the observed rainfall and the model parameters. These charcteristics will be reproduced.

o Enhancements suggested above for improved simulation of interstorms will increase the explicit correspondance between parameters and characteristics.

o There will always be charactersitics which cannot be represented explicitly by the parameters but these characteristics should, where possible, be implicit in the parameters so that they too will be reproduced. An example is annual and monthly totals of rainfall. Reproduction of the distributions of storm lengths, interstorm lengths, and rainfall amounts assures that these totals will also be reproduced.

o Streamflow simulated using stochastically generated rainfall appears to be from the same population as the observed streamflow in all tests to date. This includes a comparison of flood peaks, annual total flow, and daily flow duration.

o The procedure tested in this study appears to be the only way that long flow records can presently be developed for studying floods of long return period.

o Several results discovered in the course of this research appear to be new to stochastic modeling in hydrology.

a. The current work is apparently the only example of the use of the crossing properties of a stochastic process (the background process) to estimate parameters.

b. The discovery, albeit self-evident after some thought, of the relationship between the storm length distribution and the correlogram is also absent from the

BEST COPY AVAILABLE

literature.

c. Finally, the use of the sum of two independent random variables to transform the background process to maintain the proper distribution and at the same time reduce the autocorrelation of the generated rainfall is a significant extension of a similar technique reducing only the autocorrelation reported in Cox and Miller [1965].

BEST COPY AVAILABLE

VIII. FOLLOW ON RESEARCH

The results of this study combined with the previous work by Ott and Nasseri seem to indicate that the concept suggested in Section II offers a possible way of defining the probability of very rare floods. There is no way in which an absolute answer to the question "Is this extended frequency curve a sample from the actual population of floods for this stream?" can be obtained. Statistical tests can only say that there is only **x** percent probability that the synthetic data are **not** from the actual population of floods on the stream. A sufficiently large number of trials with similar conditions would enhance confidence in the results. Thus, tests on other streams representing a variety of climatological and hydrologic regimes seems to offer the only route to increasing confidence in or rejection of the idea. Hence, the central focus of any carry-on project should be repetition of the study reported here on other catchments.

In the course of the current study several other questions have arisen that require answers before the procedure can be considered operational. Among these are:

1. Does the conversion from analog recorders read to 0.01 inch to punched-tape recorders readable only to 0.1 inch require change in the procedure for estimating parameters for the stochastic rainfall generation? Such a change has taken place in the U. S. beginning about 1960. Obviously, values to the nearest 0.1 inch are consistent with the accuracy of rainfall measurement in general. The change however leads to a vastly increased number of 1-hr storms having 0.1 inch of rain, and a general reduction in the length of the longer storms. The procedure described in Chapter V should be tested on some 0.1-inch increment data and modified as necessary.

2. Are there regional differences in the characteristics of rainfall which reguire different

BEST COPY AVAILABLE

methods for generating the stochastic rainfall, i.e., thunderstorms, hurricanes, snow? Trials on a a number of catchments should answer this question if the selected catchments represent a sufficient variation of climatic characteristics.

3. In this study and all other trials, evaporation has been represented by a random repetition of the historic evaporation record. Can the results be improved by generating a synthetic record of potential evapotranspiration by a stochastic procedure which incorporates the cross correlation between evaporation and precipitation?

4. Can the approach be applied to catchments which experience mixed rain and snow? This would require a procedure for generating temperature data stochastically for the snow melt computations.

5. For catchments over 1000 sq. mi. it is highly desirable to use more than two raingages. What are the mathematical and computational problems involved in the stochastic analysis if one attempts to generate several rainfall traces plus evaporation and temperature?

There are clearly important questions to be answered before the concept investigated in this study is ready for operational use. The general approach required is to explore the technique on many different catchments while at the same time testing some of the specific issues discussed above. Some questions can probably be answered in trials on one catchment. Others may require replication. As the number of catchments analyzed increases, the answer to the basic question -- "Does it work?" -- will become more evident.

BEST COPY AVAILABLE

REFERENCES

Bays, C. and S. D. Durham, Improving a Poor Random Number Generator, ACM Trans. on Mathematical Software, v. 2, n. 1, pp. 59-64, Mar. 1976

Benson, M. A. Uniform Flood-Frequency Estimating Methods for Federal Agencies, Water Res. Research, v. 4, pp. 891-898, October, 1968

Bras, R. L., and I. Rodriquez-Iturbe, "Rainfall-Runoff Runoff as Spatial Stochastic Processes: Data Collection and Synthesis", Ralph M. Parsons Lab. Rept. No. 196, Mass. Inst. Tech., 1975

Bras, R. L., and I. Rodriquez-Iturbe, "Random Functions and Hydrology," Addison-Wesley, Reading, Mass., 1985

Cox, D. R. and H. D. Miller, "The Theory of Stochastic Processes," Wiley, New York, NY, 1965

Crawford, N. H., and Ray K. Linsley, "Digital Simulation in Hydrology: The Stanford Watershed Model IV", Stanford Univ. Dept. Civil Eng. Tech. Rep. No. 39, 1966

Crutcher, H. L., A Note on the Possible Misuse of the Kolmogorov-Smirnov Test, J. Applied Meteorology, v. 16, Dec. 1975

Dennis, J. E., Jr., and R. B. Schnabel, "Numerical Methods for Unconstrained Optimization and Nonlinear Equations," Prentice-Hall, Englewood Cliffs, N. J., 1983

Duckstein, L. M., M. M. Fogel, and C. C. Kisiel, A Stochastic Model of Runoff-Producing Rainfall for Summer Type Storms, Water Resources Research, v. 8, n. 2, Apr. 1972

Franz, D. D., Hourly Rainfall Synthesis for a Network of Stations, J. Hyd. Div., Proc. ASCE, v. 77, Sep., 1971

Grace, J. T., and P. S. Eagleson, "The Synthesis of Short Time Increment Rainfall Sequences," Ralph M. Parsons Lab. Rep. No. 91, Mass. Inst. Tech., 1966

Gringorten, I. I.: A plotting rule for extreme probability paper, J. Geophys. Res.,, v. 68, pp. 813-814, 1963

Hershfield, D. M.: Method for Estimating Probable Maximum Rainfall, J. Am. Water W. Assoc., v. 57, pp. 965-972, 1965

BEST COPY AVAILABLE

Johnston. J., "Econometric Methods," McGraw-Hill Book Co., New York, 1972

Kelman, J., "Stochastic Modeling of Hydrologic Intermittent Daily Processes," Hydrology Paper No. 89, Colo. State Univ., 1977

Knuth, D. E., "The Art of Computer Programming, vol. 2: Seminumerical Algorithms," Addison-Wesley, Reading Mass., 1981

Marsaglia, G., Generating a Variable from the Tail of the Normal Distribution, Technometrics, v. 6, pp. 101-102, 1964

Marsaglia, G. and T. A. Bray, A Convenient Method for Generating Normal Variables, SIAM Rev., v. 6, pp. 260-264, 1964

Morgan, B. J. T., "Elements of Simulation," Chapman and Hall,
London, 1984

Morrison, D. F., "Multivariate Statistical Methods," McGraw-Hill, New York, N. Y., 1960

Nasseri, I., "Regional Flow Frequency Analysis Using Multi-station Stochastic and Deterministic Models," Ph. D. dissertation, Dept. Civil Eng., Stanford Univ., 1972

Natural Environmental Research Council, Flood Studies Report, v. 1, p. 1.3.2, London, 1973

Ott, R. F.: "Streamflow Frequency Using Stochastically Generated Hourly Rainfall," Stanford Univ. Dept. Civil Eng. Tech. Rep. No.151, December 1971

Parzen, E., "Modern Probability Theory and Its Applications, John Wiley, New York, N. Y., 1960

Pattison, A., "Synthesis of Hourly Rainfall Data," Water Resources Research, v. 1, m. 4, pp.489-498, 1965

Richardson, C. W., "A Model of Stochastic Structure of Daily Precipitation over an Area," Hydrology Paper No. 91, Colo. State Univ., 1977

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. W. Lane,
"Applied Modelling of Hydrologic Times Series," Water Resources Pubs., Littleton, Colo., 1980

U. S. Water Resources Council, "A Uniform Technique for Dtermining Flood Frequency," Bull. 15, 1976

U. S. Water Resources Council: Guidelines for Determining Flood Flow Frequency, Hydrology Subcommittee Bull 17B, Rev. 1981

BEST COPY AVAILABLE

U. S. Weather Bureau, "Probable Maximum Precipitation in California: Interim Report", Hydrometeorological Report No. 36, October, 1961

Weibull, W.: A Statistical theory of the Strength of Materials, Ing. Vetenskapsakad. Handl Stockh. v. 151, 1939

BEST COPY AVAILABLE

APPENDIX A

Mathematical Topics

Some mathematical background is presented in this appendix to avoid unduly interrupting the presentation in the main body of Chapter 5.

Multivariate Normal Distribution

The chrarcteristics of the multivariate normal distribution are central to the stochastic rainfall model. Some of its important characteristics are given here. Let x be a column vector of variables and let it be divided into two columnar subvectors, x_1 and x_2 with at least one element in each subvector. Denote the column vector of mean values by m and the corresponding subvectors as m_1 and m_2. The covariance matrix, S is partitioned in a corresponding manner as

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

This matrix is symmetric so that $S'_{12} = S_{21}$ where the prime indicates the transpose of the matrix. Note that any symmetric postive definite matrix can be a valid covariance matrix. Morrison [1967] shows that the distribution of x_1 conditional on x_2 is multivariate normal with mean vector given by

$$m_1 + S_{12}\, S_{22}^{-1}\, (x_1 - m_2) \qquad \text{(A-2)}$$

and covariance matrix given by

$$S_{11} - S_{12}\, S_{22}^{-1}\, S'_{12} \qquad \text{(A-3)}$$

Equation A-2 and A-3 are suitable for developing the parameters of an AR(n) univariate or multivariate process. The values at the current time and at the n previous times are multivariate normal random variables. Let M_0 = lag-zero correlation matrix, M_1 = lag-1 correlation matrix, and so forth for as many lags as needed. We use correlation matrices

BEST COPY AVAILABLE

because in our case the background process is multivariate standard normal. This the correlation matrix becomes the same as the covariance matrix and the mean values are all zero. M_0 is symmetric but the other lagged matrices are not in general symmetric. For example, the lag-1 correlation matrix gives the correlations amomg all the variables that differ in time by 1 time unit. As a result the diagonal of M_1 gives the lag-1 autocorrelation for each of the AR(n) processes. The off diagonal elements give the lag-1 cross correlations between all the individual process taken pairwise. For concreteness, assume that a lag-3 multivariate process is to be represented and that the 4 correlation matrices are available (M_0, M_1, M_2, M_3). Then all we need do is define S properly to compute the parameters to use in the process. Let

$$S = \begin{bmatrix} M_0 & M_1 & M_2 & M_3 \\ M_1' & M_0 & M_1 & M_2 \\ M_2' & M_1' & M_0 & M_1 \\ M_3' & M_2' & M_1' & M_0 \end{bmatrix}$$

The vector of random variables is then partitioned in a corresponding manner. For example if each correlation matrix M_i is 2 x 2 then $x = \{x_0, x_1, x_2, x_3\}$ where the subscript for each subvector gives its lag from time zero. In this case 8 random variables are involved with the 2 variables in x_0 being conditioned by the values of the 6 variables in x_1 through x_3. The beauty of this approach is that it will work for all lags and for all cases using well defined matrix operations. It is possible to define the operations implied by Eqs. 5-2 and 5-3 explicitly if the lag is small, say 1 or 2. Even then the chance for error in implementing all the submatrix operations is great.

Sampling from the Multivariate Distribution

Let S be the conditional covariance matrix derived above and let m be the vector of conditional means. We further

BEST COPY AVAILABLE

require that S be positive definite. Let L be a lower triangular matrix (all elements above the main diagonal are zero). Then it is possible to compute L such that LL' = S. One way to do this is to use the so called Crout or square root method to factor a symmetric positive definite matrix. Let n = dimension of S. Then let

$$L_{ii} = (s_{ii} - \sum_{j=1}^{i-1} L_{ij}^2)^{0.5} \qquad i = 1,\ldots,n$$

$$L_{ki} = (s_{ki} - \sum_{j=1}^{i-1} L_{ij}\, L_{kj})/L_{ii} \qquad k = 1+1,\ldots,n$$

If S is not positive definite then the argument to the square root will become zero or negative at some point in the computations.

Once S is factored then a sample from the multivariate normal distribution with mean vector m and covariance matrix S is given by

$$x = m + Ly$$

where y is a vector of n independent standard normal random variables.

Generation of Random Numbers

To simulate a stochastic process, a source of random numbers is needed. The numbers used are generated by purely arithmetical means and are not strictly random. Methods can be devised to produce a sequence of numbers which pass a series of intuitively satisfying tests. Knuth [1981] and Morgan [1984] among others discuss these methods and tests at length.

A multiplicative congruential generator was used with a multiplier of 314159269 with all integers being taken modulo 2147483647 because the computer used for this study utilized the two's complement representation of integers in a 32-bit word. Knuth [1981] reported the results of the spectral test

BEST COPY AVAILABLE

with this multiplier and it had good characteristics. The same multiplier was used by Franz [1970].

The random numbers produced by this multiplier were further randomized using a shuffling technique reported by Bays and Durham [1976]. A table of size 256 was used for speed of execution in Fortran. Knuth [1981] reports that this method can be used to improve any random number generator with only modest additional cost.

The multiplicative congruential method produces random integers and, if desired, a random real number uniformly distributed between 0 and 1. However, distributions other than the uniform can be derived from these uniformly distributed numbers.

The standard normal random numbers required were developed using a composition approach reported by Marsaglia and Bray [1964]. This method represents a good balance between the speed and ease of programming in a language like Fortran. The truncated normal variables required in computed the simulated rainfall were computed using a method reported by Marsaglia [1967]. Let u and v be two independent uniform random variables such that $0<u, v<1$. Set

$$x = (a_2 - 2\ln v)^{0.5}$$

where ln = natural logarithm. Then if $vx < a$, accept x as a valid value. Otherwise reject u and v and start over. Then x is a standard normal random variable condition such that $x>a>0$. This technique uses what is called the rejection method and is discussed at length in Knuth [1981] and Morgan [1984].

BEST COPY AVAILABLE

NRC FORM 335 (11-81)

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1. REPORT NUMBER *(Assigned by DDC)*: NUREG/CP-4496

4. TITLE AND SUBTITLE *(Add Volume No., if appropriate)*: A System for Generating Long Streamflow Records for Study of Floods of Long Return Period

2. *(Leave blank)*

3. RECIPIENT'S ACCESSION NO.

7. AUTHOR(S): R. K. Linsley, B. A. Kraeger, D. D. Franz

5. DATE REPORT COMPLETED: MONTH April, YEAR 1985

9. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*:
Linsley, Kraeger Associates LTD
527 Bayview Drive
Aptos, CA 95003

DATE REPORT ISSUED: MONTH February, YEAR 1986

6. *(Leave blank)*

8. *(Leave blank)*

12. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*:
Division of Risk Analysis and Operations
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

10. PROJECT/TASK/WORK UNIT NO.

11. FIN NO.: D1562

13. TYPE OF REPORT: Informal

PERIOD COVERED *(Inclusive dates)*

15. SUPPLEMENTARY NOTES

14. *(Leave blank)*

16. ABSTRACT *(200 words or less)*

A stochastic model for the multi-station generation of hourly rainfall has been developed and tested. The model uses a multivariate-normal background process which is continuous. The process is truncated to yield intermittent output as rainfall. The model results showed excellent agreement between a 1000-yr series and observed rainfall. The synthetic rainfall was used with a hydrologic simulation model to develop a synthetic flow record of 1000-yrs. For a portion of the of the Russian River near Cloverdale, CA tests of these data indicate that they represent the same population used in calibration of the two models.

17. KEY WORDS AND DOCUMENT ANALYSIS

17a. DESCRIPTORS

17b. IDENTIFIERS/OPEN-ENDED TERMS

18. AVAILABILITY STATEMENT: Unlimited

19. SECURITY CLASS *(This report)*: Unclassified

21. NO. OF PAGES

20. SECURITY CLASS *(This page)*: Unclassified

22. PRICE: S

NRC FORM 335 (11-81)

BEST COPY AVAILABLE